DOCUMENT RESUME

ED 199 277 TM 810 182

AUTHOR Douglass, James B.
TITLE A Comparison of Item Response Theory Models for Use in a Classroom Examination System. Promising Applications of Latent Trait Models and Evidence for Their Validity.
PUB DATE Apr 81
NOTE 111p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Academic Achievement; *Comparative Analysis; Difficulty Level; Equated Scores; Higher Education; *Latent Trait Theory; Mathematical Formulas; Sampling; *Testing; Test Items; True Scores
IDENTIFIERS Item Calibration; Rasch Model; Three Parameter Model; Two Parameter Model

ABSTRACT

Methods and results relevant to the introduction of item characteristic curve (ICC) models into classroom achievement testing are provided. The overall objective was to compare several common ICC models for item calibration and test equating in a classroom examination system. Parameters for the one-, two- and three-parameter logistic ICC models were estimated for 100-item final examinations taken over the course of four academic terms by 594 to 1082 examinees in an introductory college-level communications course. The three-parameter model provided unacceptable lower asymptotes for this data set and was given no further consideration. Rasch calibration and two methods of two-parameter model calibration were investigated and their estimated calibration constants were used with a true-score method to equate the examinations to a common scale. The Rasch model equatings were very consistent across all sample sizes investigated. Both of the two-parameter methods provided equatings that were less consistent. There was, however, evidence that Rasch item difficulties were less stable across examinations and academic terms than they were across different examinee samples taken from a single test administration. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED199277

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A COMPARISON OF ITEM RESPONSE THEORY MODELS FOR USE IN A CLASSROOM EXAMINATION SYSTEM

By

James B. Douglass
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J.B. Douglass

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented in the AERA/NCME Symposium

Promising Applications of Latent Trait Models
and Evidence for Their Validity

TM 810 182

Paper presented at the AERA/NCME Annual Meeting in Los Angeles, CA, April 13-17, 1981. Although many people contributed to this paper, the author wishes particularly to thank Dr. Robert L. Ebel, Dr. LeRoy A. Olson and Debbie Bennett. The research upon which this paper is based was done while the author was with the Learning and Evaluation Service, Michigan State University.

ABSTRACT

Item characteristic curve (ICC) models offer the potential of producing a measurement of a person's ability or achievement that does not depend on the particular sample of items in a test or on the particular examinees upon which test analysis is based. The present study is designed to provide methods and results relevant to the introduction of item characteristic curve models into classroom achievement testing. The overall objective is to compare several common ICC models for item calibration and test equating in a classroom examination system.

Parameters for the one-, two- and three-parameter logistic ICC models were estimated for 100-item final examinations taken over the course of four academic terms by 594 to 1082 examinees in an introductory college-level Communications course. The three-parameter model provided unacceptable lower asymptotes for this data set and was given no further consideration. One of the four tests was chosen to provide a base scale to which items from the other examinations were calibrated using the one- and two-parameter models. There were 43 to 53 items in common between the base scale examination and each of the other three. Random samples of 200, 600 and 800 examinees were selected from a total sample of 1082 examinees to investigate the impact of examinee sample size upon item

calibration and test equating. Rasch calibration and two methods of two-parameter model calibration were investigated. One method of two-parameter model calibration was based on item difficulty means and standard deviations, and the other was based on mean item difficulties and ratios of item discriminations. Calibrations based on the sample of 200 examinees and all 43 items in the calibration section were quite divergent across methods and also diverged from the calibrations based on the other samples.

The calibration constants estimated from the Rasch model and the two methods using the two-parameter model were used with a true-score method to equate the examinations to a common scale. The Rasch model equatings were very consistent across all sample sizes investigated. Both of the two-parameter methods provided equatings that were less consistent.

Consistency of equating provides only indirect evidence of equating accuracy. To provide a criterion for equating, a test was equated to itself using three separate sets of examinee samples selected from the 1082 examinees in the largest total sample. The three sets represented random, high-low and very high-very low examinee ability differences. Although none of the three methods was uniformly best, the Rasch model provided the most acceptable equating in general.

The stability of Rasch calibrating links and hence ability estimation based on 16 class sections ranging in size from 49 to 66 examinees was investigated for a 43 item test taken from a total test of 100 items. Calibration links were calculated for subsets of the 43 items ranging in size from 7 to 37 in increments of five items. Even with 37 items in the calibrating link, it was found that there was an average bias in ability estimation equivalent to 1/5 of a standard deviation of the ability estimates based on all 43 items.

For this testing system, it is concluded that the Rasch model provided test equatings as good as or better than were provided by either two-parameter method. There was, however, evidence that Rasch item difficulties were less stable across examinations and academic terms than they were across different examinee samples taken from a single test administration. It is recommended that the long-term stability of parameter estimates, and the scales based upon them, be systematically monitored and studied in testing systems using ICC models.

Copyright by

James Bruce Douglass

1980

TABLE OF CONTENTS

Chapter		Page
I.	STATEMENT OF THE PROBLEM	1
	Objectives	3
	Historical Perspective	4
II.	COMPARING MODELS	8
	An Overview of Model Comparisons	8
	A Process for Comparing Models	12
	Previous Model Comparisons	14
III.	METHOD	24
	Sample	24
	Models	26
	Estimation of Model Parameters	28
	Scaling Model Parameter Estimates	28
	Equating	32
	Procedure	33
IV.	RESULTS	36
	Descriptive Data	36
	Parameter Estimation	38
	Calibration	42
	Rasch Model	42
	Two-parameter Model	46
	Test Equating and Ability Scaling	48
	Study 1	48
	Study 2	66
	Study 3	66
V.	SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	78
	Summary	78
	Conclusions and Discussion	80
	Parameter Estimation	80
	Calibration	81
	Test Equating and Ability Estimation	82
	Recommendations	84
APPENDIX		86
LIST OF REFERENCES		96

LIST OF TABLES

Table		Page
4.1	Descriptive Data for Study 1	37
4.2	Descriptive Data for Study 2	37
4.3	Descriptive Data for Study 3	39
4.4	Two-parameter Item Difficulty ($\hat{b}$) Statistics for Study 1	41
4.5	Two-parameter Item Difficulty ($\hat{b}$) Statistics for Study 2	41
4.6	Rasch Model Calibration for Study 1.	43
4.7	Rasch Model Calibration for Study 2.	45
4.8	Two-parameter Model Calibration for Study 1.	47
4.9	Two-parameter Model Calibration for Study 2.	49
4.10	Equating Inconsistency Across Models	63
4.11	Equating Inconsistency Within Models	64
4.12	Equating Error Across Models	73
4.13	Rasch Ability Estimates for Study 3.	75
4.14	Rasch Ability Estimate Error Due to Calibration for Study 3	76
A1	Difficulties ($\hat{b}$) of Items Removed Before Calibration	95

LIST OF FIGURES

Figure		Page
4.1	Spring 1979 Equating	51
4.2	Fall 1978 Equating	52
4.3	Winter 1979 Equating	53
4.4	Fall 1979 Equating	54
4.5	Fall 1979 Sample of 200 Equating	55
4.6	Fall 1979 Sample of 600 Equating	56
4.7	Fall 1979 Sample of 800 Equating	57
4.8	Rasch Equating Fall 1979 Samples	59
4.9	B Method Equating Fall 1979 Samples	60
4.10	A Method Equating Fall 1979 Samples	61
4.11	Rasch Random Halves	67
4.12	Rasch Low to High Half	68
4.13	Rasch Very Low to Very High Half	69
4.14	2 Par Random Halves	70
4.15	2 Par Low to High Half	71
4.16	2 Par Very Low to Very High Half	72
A1	F78 and S79 Diff Est	86
A2	W79 and S79 Diff Est	87
A3	F79 and S79 Diff Est	88
A4	F79 (N = 200) and S79 Diff Est	89
A5	F79 (N = 600) and S79 Diff Est	90

Figure		Page
A6	F79 (N = 800) and S79 Diff Est	91
A7	F79 Random Halves Diff Est	92
A8	F79 Low-High Halves Diff Est	93
A9	F79 Very Low-High Halves Diff Est	94

CHAPTER I

STATEMENT OF THE PROBLEM

Item characteristic curve (ICC) models, also called latent trait or item response models, have drawn increasing interest in recent years from both psychometric researchers and practitioners (Hambleton and Cook, 1977; Wright, 1977; and Hambleton, et.al., 1978). The primary appeal of ICC models lies in their potential to solve the problem of deriving comparable scores from different tests given to different people. For example, in Communication 100, an introductory course at Michigan State University, different instructors give different mid-term examinations to different sections. If one section receives lower scores than another, does it mean that the students in that section have less knowledge or that they received a harder test? It would be highly desirable to devise a common scale to which scores from each mid-term could be referenced. Although a common final examination is administered each term, the finals differ across terms. Again a common reference scale would be useful. Students' grades should not depend upon the difficulty of the particular tests they take. Nor should their grades depend on the grades of other students in their particular section. If the potential of ICC models to provide item statistics that are independent of particular examinees and ability scores that are independent of particular items can be realized in the classroom, then score comparability can be achieved in a classroom testing system.

Relatively little ICC research has focused on achievement testing and even less has been addressed to classroom examination. The present study is one of the first to systematically apply ICC models to a classroom examination system. There are two potential problems in applying ICC models to classroom examinations. The first problem is that if different teachers teach the same material in different ways, different factors are likely to result (Lord, 1980). In other words, teachers and teaching methods can affect the factor structure of examinations. Although a single factor is only a sufficient but not necessary condition for the ICC model assumption of unidimensionality (Lord and Novick, 1968), classroom achievement tests are likely to be less unidimensional than aptitude tests. Furthermore the dimensionality of achievement tests could vary from term to term and class to class depending on differences in instructors and instructional methods.

The second problem has been recognized by Bejar, Weiss and Kingsbury (1977). They note that aptitude test item pools can typically be constructed and refined to meet the unidimensionality assumption quite well. However, a classroom achievement test must fairly represent the entire course content. There is much less freedom to delete areas of content in classroom tests than there is in aptitude tests.

Fortunately the question of interest is not whether the assumptions of ICC models are met. In practice, assumptions are never completely met. The practical questions are:

1. To what extent are the assumptions of the models met?
2. What effects do violations of the assumptions have on the desired properties of the models?

Robustness of the models to violations of their assumptions is a central concern.

There is a problem resulting from the robustness issue which has been largely ignored in the ICC literature. As an illustration of the problem, consider Wright's (1977, p. 113) procedure for designing a "best test." His procedure involves only psychometric properties of the items in the test and guesses about the mean and dispersion of the abilities of persons to be tested. There is no mention of item content. Of course, if an item pool is truly unidimensional, then the content of any particular item may be irrelevant. But a model may give consistent and valid results either because its assumptions are met or because it is robust with respect to violations of its assumptions. A model may show evidence of fit despite a lack of unidimensionality. Therefore the content validity of a test is still a relevant concern in test development. The use of latent trait models with classroom achievement tests does not relieve the test constructor of the task of developing a test with a balanced representation of course content.

OBJECTIVES

The purpose of this study is to determine whether item characteristic curve models can be successfully applied to a classroom achievement testing system. The overall objective is to compare the one-, two-, and three-parameter logistic item characteristic curve models for item calibration and test equating in a large multi-section college course. The overall objective will be accomplished through the following specific objectives:

i. To obtain and examine parameter estimates for all three models.

ii. To calibrate items to a common scale using the one- and two-parameter models. Calibration constants will be estimated

using different numbers of examinees. This will result in different calibrations which will be examined for consistency.

iii. To equate tests to a common scale using the one- and two-parameter models. Equating will be accomplished using the different calibrations mentioned in ii. This will result in test equatings that will be compared for consistency within and across models.

iv. To equate a test to itself using the one- and two-parameter models. The equating will be accomplished using samples of examinees that have different average levels of ability. The accuracy of equating will be compared for the various models.

v. To obtain ability estimates from different examinee sections using the one-parameter model on a set of items common to all sections. Ability estimates will be compared for each class section assuming that different numbers of items within the common set of items were previously calibrated. This simulates the situation of an instructor adding some of his/her own items to a test composed of previously calibrated items.

HISTORICAL PERSPECTIVE

Lord (1968, 1980) states that Lawley (1943) was the first to present a coherent test theory based on item characteristic curves. Lawley assumed that there was no guessing, ability was normally distributed and that item intercorrelations were equal. Lazarsfeld (1950) presented a

more general theory and may have been the first to use the term "latent traits" (Hambleton, et.al., 1978). However, Lord's (1952, 1953a, 1953b) initial work with the ogive model is considered by many to be the real beginning of latent trait theory applied to mental tests. Lord himself does not use the term "latent trait theory". He originally spoke of item characteristic curve theory and now uses the term "item response theory" (Lord, in press). In Lord's early work with the normal ogive, he assumed no guessing but unlike Lawley, did not assume a normal distribution of ability or equal item intercorrelations.

Lord (1980) stopped work for ten years on item characteristic curve models for three reasons:

1. Maximum likelihood estimators for all parameters were computationally impractical.

2. His early work assumed responses were either correct or incorrect. There was no provision for omitted or not reached items.

3. It was not clear that one can assume that the probability of correct response to an item can be represented as a normal ogive function of ability. In particular, if item-test regressions are not monotonic it would be difficult to fit any simple mathematical model (Lord, 1965).

It had been argued that item-test regressions might start at chance level, then decline before rising because the worst examinees would answer randomly while slightly better students would choose plausible distractors. Another argument against monotonicity was that the very best examinees might do slightly worse than somewhat less able examinees. The contention was that the very best students might consider factors beyond the scope of the item and thus answer incorrectly. Lord (1965) investigated the monotonicity of item-test regressions using 103,275 examinees and 90 verbal and 60 mathematical items from the May 2, 1964

administration of the Scholastic Aptitude Test. He found that except for chance fluctuations, the item-test regressions were monotonic throughout the range of test scores.

The computational difficulties were greatly reduced by Lord's (1968) adoption of Birnbaum's (1968) three-parameter logistic model. This coupled with advances in high speed computers made parameter estimation feasible. The three-parameter model, unlike Lord's earlier two-parameter normal ogive model, assumes that guessing can occur. Lord (1974) also modified the likelihood function to handle omitted and not reached responses as well as correct and incorrect responses. The major impediments to progress on item characteristic curve models were thus removed.

Meanwhile an independent line of development in latent trait models was occurring. In Denmark, Rasch (1960) was working on several different measurement models based on properties he felt mental measurements should possess. One of these models has the mathematical form of a one-parameter logistic item characteristic curve model. This has become known as the Rasch model. Wright (1968) introduced the Rasch model in the United States and has since become it's leading proponent.

Nineteen sixty-eight was a very significant year for item characteristic curve theory. In that year, a rigorous mathematical presentation of the three-parameter logistic model (Birnbaum, 1968) was published. Lord (1968) publicly adopted Birnbaum's model and Wright (1968) introduced the Rasch model in the United States. The next major event occurred in 1977 with the publication of the summer issue of the Journal of Educational Measurement. The entire issue consisted of six invited papers on latent trait models. The issue represented the first attempt to present

current developments in latent trait theory at a level understandable to the average measurement specialist. A year later a review of the latent trait literature appeared (Hambleton, et.al., 1978). Perhaps the ultimate evidence of the acceptance of latent trait theory was the publication of two textbooks in the area (Wright & Stone, 1979 and Lord, in press). The popularization of latent trait theory combined with the availability of computer programs (Wood, Wingersky & Lord, 1976 and Wright, Mead & Bell, 1979) has led to a proliferation of research in the area.

CHAPTER II

COMPARING MODELS

The first section of this chapter presents a conceptual framework useful for analyzing model comparisons and studies of fit. A process by which models can be compared for the solution of practical testing problems is developed in the second section. The final section contains a review of previous latent trait model comparisons.

AN OVERVIEW OF MODEL COMPARISONS

Model comparisons and tests of fit can be categorized by the antecedent and consequent conditions investigated, by the anticipated generalizability of results and by the degree of artificiality in the situation or data. In their study of the Rasch model, Rentz and Bashaw (1975) distinguish between assumptions of the model, antecedent conditions derivable from the assumptions and consequent conditions derivable from specific objectivity. They associate tests of item fit with antecedent conditions such as unidimensionality, equal item discriminations and the absence of guessing. Others (e.g. Hambleton and Cook, 1977) classify these conditions as assumptions. For Rentz and Bashaw the consequent conditions of primary interest are the invariance of item parameter and ability estimates. Antecedent conditions are closely associated with test construction and consequent conditions are associated with analyzing existing tests.

Antecedent and consequent conditions can be further subdivided (cf. Douglass, 1979). Antecedent conditions can be categorized as assumptions or derivations from assumptions, as operational definitions of models defined by computer programs or estimation procedures, or as situational or design conditions. Assumptions of latent trait models have been discussed in the literature in some detail (Hambleton & Cook, 1977 and Lord & Novick, 1968). The particular models used in this study are discussed in Chapter III.

Much less thought seems to have been given to how the models are operationalized. Different computer programs estimate the parameters of the models differently. For example, the LOGIST program is capable of estimating parameters of the one-, two- or three-parameter logistic model. The program uses item response data categorized as correct, incorrect, omitted or not reached. On the other hand, BICAL, which estimates parameters for the one-parameter or Rasch logistic model, only accepts item responses of correct or incorrect. Although LOGIST and BICAL both estimate parameters of the one-parameter logistic model, somewhat different estimates could result if there are a significant number of omitted or not reached items in a test.

The procedure used to estimate parameters can make a real difference in results obtained. Using simulated data, Ree (1979) compared four procedures for obtaining estimates for the three-parameter model. He found that none of the four procedures, LOGIST, ANCILLES, OGIVIA, and a transformation of the item-raw score biserial, was uniformly superior. The best procedure depended on the distribution of ability in the calibration sample, the intended use of the parameter estimates and the computer resources available.

A third type of antecedent condition is a situation or design condition. These conditions can reflect contraints in a testing situation of interest or they may be manipulated by the researcher or practitioner. Examples include shapes and ranges of ability distributions, sample sizes of examinees and test items, test content and item types. Situational conditions can affect the usefulness of models by affecting either the assumptions of the models or the goodness of the estimation procedures.

Consequent conditions can be classified by the degree to which they are related to some problem of interest. At one extreme are consequences predicted from the model that have little or no direct relation to any particular practical problem of interest. Some statistical tests of fit are based on this type of consequence. These tests of fit are of very limited usefulness in determining how well a model will work in any given practical situation of interest. Probabilities associated with tests of fit are highly dependent on sample size. Furthermore, there is rarely a known relationship between a fit statistic and a practical criterion of interest. However, this type of fit statistic may have some utility for comparing degrees of fit across models. Of course, fit statistics are quite appropriate when applied to antecedent conditions in an attempt to better meet assumptions. For example, tests of item fit can be very useful in building a test that is more likely to result in desired consequences.

At the other extreme are consequent conditions that are of direct interest for a problem under investigation. For example, if one is equating tests, consistency of equatings based on different examinee samples is a consequent condition of latent trait models that is of

direct interest. The comparison of a predicted consequence of a model with an actual consequence may contain little or much useful information.

Studies of model fit and model comparison can also be categorized by the type of data and situations investigated. Data can be real or simulated and situations can be realistic or contrived. Generally the trade-off is between realism and control. With simulated data the parameters of the model are known. Ability and item distributions can be completely specified. Unfortunately the parameters and distribution may not accurately reflect real data sets. On the other hand, real data comes from some actual situation, but the true parameters are unknown. Of course, it is possible to have realistic simulations or unrealistic real data sets. Simulations can include many relevant variables and real data sets may come from a very contrived situation or from a very poor sampling plan.

When considering model comparisons and tests of fit, it can be helpful to consider how general the results are intended to be and how general they are in fact. Cronbach (1975) has argued that the search for generic laws and grand theories, so successful in the natural sciences, is doomed to failure in the behavioral sciences. He argues that the problem is not that human events are unlawful, but that the times are a relevant variable, so generalizations can not be stored up for assembly into a causal network. Cronbach feels that behavioral science will be better served if generalizations from research studies are treated as working hypotheses rather than as conclusions.

Cronbach's views have several implications for those attempting to apply models to human behavior, including item response behavior. One implication is that a model may work quite well in one situation and

not at all well in another. It is necessary to be alert to local conditions that may be influencing one's results. A second implication is that it is always necessary to check to see if a model is working satisfactorily in any given situation of interest. Furthermore, since results may change over time, it is necessary to constantly check to see that a model continues to work satisfactorily. If science is thought of as a discipline where laws are determined and theories built and engineering is thought of as the application of those laws and theories to practical situations, it may be that in the past those in the behavioral sciences have done too much poor science and too little good engineering. Cronbach (1975, p. 126) argues that we can realistically expect to achieve two goals in the behavioral sciences, "... to assess local events accurately, to improve short-run control ... (and) to develop explanatory concepts, concepts that will help people use their heads."

A PROCESS FOR COMPARING MODELS

A process developed by Douglass (1979) for comparing models will be outlined in this section and serves as the basis of the research strategy for this study. The process includes consideration of the three types of antecedent conditions mentioned previously, effects of antecedent conditions on relevant consequent conditions and focuses on the particular practical problem of interest in a situational context. The process can be thought of as an elaboration and extension of a strategy suggested by Lord and Novick (1968) for investigating the practical utility of a model.

Lord and Novick (1968, p. 383) suggest the following four steps for investigating the practical utility of a model:

1. Estimate the parameters of the model, assuming it is true ...
2. Predict various observable results from the model, using the estimated parameters.
3. Consider whether the discrepancies between predicted results and actual results are small enough for the model to be useful ("effectively valid") for whatever practical application the investigator has in mind.
4. If in step 3 the discrepancies were considered too large, then it may be useful to compare them with the discrepancies to be expected from sampling fluctuations.

Douglass (1979) extends this procedure to include comparisons of different models for consistency, validity, and utility. The basic steps Douglass suggests are:

1. Identify the models to be compared.
2. Obtain different sets of parameter estimates for each model, assuming each is a true representation of reality. These estimates should be obtained in the actual situation where the selected model will be used and should reflect any sources of error that will be operating in the practical setting. In addition to naturally occurring variation in antecedent conditions, variation can be introduced artificially to simulate the practical situation. For example, different samples can be investigated by splitting a large sample into sub-samples. This will result in discrepant estimates for each model.
3. Observe the effects of the discrepant estimates for each model on each consequent condition of interest.
4. If any given model does not give sufficiently consistent results, then the antecedent conditions should be investigated. Sources of inconsistency include sampling error or bias, estimation procedure error and violation of model assumptions.

5. Compare the results of the models which give consistent results or can be made to give consistent results to one another.
6. Investigate the patterns of convergence and divergence both between and within models along with any available criteria to aid in model selection.

PREVIOUS MODEL COMPARISONS

One of the first comparisons between latent trait models was made by Hambleton and Traub (1973) using data from two verbal tests and one mathematics aptitude test. Hambleton and Traub compared observed and expected test score distributions for the one- and two-parameter logistic models. For the one-parameter model, a person's test score was simply the sum of all items answered correctly. For the two-parameter model, the test score was the sum of the products of each correct item and its ICC discrimination estimate. Comparisons between the one- and two-parameter model were made using a chi-square test of fit applied to the observed and expected distributions of test scores. The comparison of observed and expected distributions of test scores is a good example of using a consequent condition that is predicted from a model but that is not directly related to any practical problem of interest. Of course, the limitation of such a comparison is that it only provides information on which model appears to be generally better. There is no way to determine how much better one model will be than another for any given practical testing problem.

Hambleton and Traub assumed that ability was normally distributed in their study. This allowed them to estimate item characteristic curve difficulties and discriminations using simple functions of classical item difficulties and item-test biserial correlations. The obvious

advantage of assuming a normal distribution of ability is that estimation is very inexpensive and does not require sophisticated computer programs such as LOGIST or BICAL. The disadvantage is that an unnecessary and restrictive assumption is added to the models.

Hambleton and Traub found that the two-parameter model predicted test performance better than the one-parameter model for each of the three sets of test data. The greatest difference was for a short test of 20 items that had a fairly large range of item discriminations (.74).

A recent study by Hutton (1980) is an extension of the work of Hambleton and Traub (1973). She examined the relationship between dimensionality, the range of item discriminations and guessing on distributions of observed and expected number-right scores for the one- and three-parameter logistic models. She compared the observed and expected distributions graphically and with chi-square and Kolmogorov-Smirnov tests of fit.

Advantages of the Hutton study over the earlier Hambleton and Traub study include a sample of 25 tests, an attempt to relate dimensionality and range of item discriminations to degree of fit and the use of LOGIST for parameter estimation, thereby making no assumptions about the underlying distribution of ability.

The degree of unidimensionality was represented by the ratio of eigenvalues between the first and second factors of a principal components factoring of the matrix of tetrachoric correlations for each test. The measure of spread of item discriminations for each test was the percentage of point-biserial item-test correlations falling within a .10 band around the mean of the point-biserials for that test. It proved impossible to find an acceptable estimate of guessing that was independent of the latent trait models.

Hutton found that the three-parameter model fit slightly better on the average than the one-parameter model using the Kolmogorov-Smirnov (K-S) test of fit. Correlations between degree of unidimensionality and K-S fit, ranging from -.555 to -.468, were significant for both the one- and three-parameter models at the .05 significance level. Correlations between degree of equality of item discrimination and K-S fit, ranging from -.238 and -.158 were not significant at the .05 level for either model. Hutton concluded that for estimating ability the one-parameter model produces estimates practically as good as those of the three-parameter model.

Several cautions are in order. First, Hutton used number-right score as a criterion. Number-right score is a sufficient statistic for estimating ability under the one-parameter model but not under the three-parameter model. An appropriately weighted score as a criterion for the three-parameter model probably would have made the model look better. Second, as Hutton observed, similarity of results for ability estimation does not imply similarity of results for other applications such as test development, test equating or investigation of item bias.

Reckase (1979) also investigated the relationship between dimensionality and fit of the one- and three-parameter logistic models. He used five real data sets and five data sets simulated to have particular factor structures. The study was designed to answer three questions (p. 226):

> (1) What component of the tests is being measured by the two models? (2) Does the size of the first factor control the estimation of the parameters of the two models? (3) What is the relationship between the factor analysis, latent trait and item analysis parameters?

He found that in the case where a test is composed of several independent factors, an uncommon situation in practice, the three-parameter model picks just one factor and differentiates among ability levels on that factor ignoring the other factors. The one-parameter model, on the other hand, estimates ability based on the sum of the independent factors. In the common situation where there is a large first factor in a test and several smaller factors, both methods measure primarily the first factor to the same extent.

The size of the first factor was found to affect parameter estimates. Reckase obtained good ability estimates when the first factor accounted for less than 10 percent of the test variance, but found that stable item calibration required that at least 20 percent of the test variance be accounted for by the first factor.

A factor analysis of the correlations between item statistics from factor analysis, latent trait analysis and traditional item analysis resulted in three factors which Reckase called difficulty, discrimination and guessing. He found that the one- and three-parameter difficulty estimates, traditional difficulty and the three-parameter discrimination all loaded on the difficulty factor. The undesirable relationship between latent trait difficulty and discrimination was also found by Lord (1975). The three-parameter discrimination also loaded on the discrimination factor, as did the traditional point-biserial. The one-parameter chi-square item fit statistic loaded with the three-parameter "guessing" parameter estimate. Guessing, not multidimensionality or item discrimination, seemed to most greatly affect the one-parameter tests of fit.

Work like that of Reckase, where there is a systematic attempt to relate consequent conditions to variations in antecedent conditions, is clearly useful. Although there is no assurance that his results will generalize to other tests and testing situations, his study does give tentative relationships and information which can serve to guide others.

The three studies just discussed considered consequent conditions that, although predicted from the models, were not directly related to particular practical testing problems of interest. The next three studies all attempt to compare models for consequent conditions of direct interest for a practical testing problem.

Hambleton and Cook (1978) compare the one-, two- and three-parameter models for rank ordering examinees. The correct rank ordering of examinees is of course relevant in any norm-referenced testing situation. They simulated data, using a three-parameter model, to study the effects of variation in range of item discrimination parameters (0, .62 and 1.24), the average value of lower asymptotes of the item characteristic curves (0 and .25), test length (20 and 40 items), and the shape of the ability distribution (uniform and normal) upon the rank ordering of examinees. All tests were constructed to be unidimensional.

Hambleton and Cook found that the three-parameter model was better at ranking examinees in the lower half of the ability distribution than the one-parameter or number-right model. Correlations of ability estimates for 20 item tests were approximately .08 higher for the three-parameter model. The difference went from approximately .75 to .83 for the uniform distribution of ability and .65 to .73 for the normal distribution of ability. For 40 item tests, correlations were approximately .03 higher for the three-parameter model compared to the one-parameter

model or number-right model. The averages were approximately .90 compared to .93 for the uniform distribution and .80 compared to .83 for the normal distribution of ability. Except for lower ability examinees, the number-right score did just about as good a job of ranking examinees as the more complicated ability estimates.

One caveat should be noted here. The rank ordering of a set of examinees on the basis of a set of test items is not a very strong use of latent trait ability estimates. Applications requiring better estimates, for example item calibration, test equating or tailored testing (McKinley & Reckase, 1980) show greater differences.

Douglass, Khavari and Farber (1979) compared classical item analysis procedures with a method based on the Rasch model for constructing a scale for assessing alcoholism. Items were selected from a 49 item total scale. The traditional method consisted of choosing items with item-total correlations greater than .20, resulting in a scale with 29 items. The Rasch selection procedure consisted of rejecting those items that had item chi-square fit probabilities less than .001. Thirty-six items remained in the Rasch scale from the original 49.

The two procedures resulted in somewhat different items being selected for each scale. Twenty-four items were common to both scales. The traditional scale contained five unique items and the Rasch scale contained 12 unique items. Despite the differences in items, the correlations between the two scales with the total scale were not significantly different. The correlations between the two scales with an outside criterion were also not significantly different.

There are several methodological problems that limit the generalizability of the results. First, the limited item pool, 49 items, necessitated a great deal of overlap in items selected by the two methods.

The overlap would tend mask differences. Secondly, the selection of a cut-off level for the item chi-square test of fit was arbitrary and is subject to fluctuation depending on sample size. A better method would be to select the best 30 items using the Rasch method and the best 30 using a traditional method with both selections from a larger item pool. Third, the correlation between the traditional scale and the total scale should be higher, other conditions being equal, since the items were selected because they had relatively high correlations with the total scale. But four, other conditions were not equal, since the Rasch scale contains more items from the original scale, 36 of 49, compared to 29 of 49 for the traditional method. In fact, the number of items made the greater difference, the correlations for the Rasch and traditional scales with the total scale were .914 and .881, respectively. The study is admirable for the emphasis on a practical consequent condition of interest but not for its methodology.

In the most massive comparative study of test equating models ever attempted, Marco, Peterson and Stewart (1980) compare nine linear observed-score models, two curvilinear observed-score models, nine linear true-score models ard two curvilinear true-score models. The two curvilinear true-score models are based on the one- and three-parameter logistic item characteristic curve models. The LOGIST computer program was used to obtain parameter estimates for both logistic models. The anchor-test method was used as the design method for all test equatings. The April 1975 and November 1975 administrations of the verbal portions of the Scholastic Aptitude Test (SAT) provided the test data for the study. Equating conditions were varied by having anchor tests internal or external to a test and similar or dissimilar in context, difficulty and length to the test being equated. The consequent condition investigated was the

goodness of test equating. This was investigated by two methods. One method was to equate a test to itself. The conversions from raw to scaled scores should be the same in this case. The other method, where two different tests were equated, was to use the most ideal equating method possible to provide a criterion. This was accomplished by equating the tests based on a single sample of 4,731 examinees who had scores on both tests being equated. Equipercentile equatings using the three-parameter logistic model and observed scores provided two criteria for the second method.

Total error in equating was defined as the standardized weighted mean square difference between the estimated criterion score and the criterion score. The total error is equal to the variance of the difference plus the squared bias. Algebraically,

$$\sum_j f_j d_j^2 / n s_t^2 = \sum_j f_j (d_j - \bar{d})^2 / n s_t^2 + \bar{d}^2 / s_t^2 \qquad (2.1)$$

where $d_j = t_j' - t_j$

t_j' is the estimated criterion score for raw score X_j,

t_j is the criterion score for X_j,

$\bar{d} = \sum_j f_j d_j / n$,

s_t is the standard deviation of the criterion scores t_j,

f_j is the frequency of X_j,

$n = \sum_j f_j$

and the summation is over the range of X for which extrapolation was not necessary for any of the models studied. The multiplication by f_j serves to weight the error more where more scores actually occur. Division by the variance of the criterion score, s_t^2, standardizes the error terms.

The variance of the difference between the estimated criterion score and the criterion score is a measure of the inconsistency between the estimated and actual criterion. The squared bias is the mean difference between the estimated and actual criterion squared and reflects systematic error. Marco, Peterson and Stewart plot the errors of each equating method for each condition studied on graphs with bias on the horizontal axis, the standard deviation of the difference between the estimated and actual criterion on the vertical axis and total error on parabolic lines. It is thereby easy to compare not only total error for the different methods, but also the two types of error.

Several conclusions drawn by Marco, Peterson and Stewart are particularly interesting.

1. Types of examinee samples have relatively little and unsystematic effect on the quality of equating results if the anchor test is similar in content and difficulty to the total test.
2. When a test is equated to a test like itself using an easy or hard anchor and random examinee samples, all of the models have small total error. But when dissimilar examinee samples are used, the one- and three-parameter ICC models are clearly superior.
3. When tests differ considerably in difficulty, linear models and the one-parameter ICC model have very large total errors.
4. The three-parameter ICC model gives better results than any of the other models under extreme or unusual conditions; for example, tests of different difficulty

given to dissimilar examinee samples. This result is of course consistent with the invariance property of ICC model parameter estimates.

As Marco, Peterson and Stewart point out, the strength of their study is the comprehensiveness of the models studied. However, all comparisons were made using verbal items from the Scholastic Aptitude Test. The SAT verbal items are relatively homogeneous and difficult for current examinees. These are conditions which favor meeting the ICC model assumption of unidimensionality but not the one-parameter ICC model assumption of zero lower asymptotes. Under these conditions, the three-parameter ICC model should work fairly well regardless of item and examinee samples. The one-parameter ICC model should work reasonably well if the lower asymptote assumption is not challenged by the data. In other words, if item difficulties and examinee abilities are reasonably similar, little guessing will occur and problems with the one-parameter model will not become apparent. On the other hand, tests of different difficulty will differently encourage guessing, causing poor results from the one-parameter ICC model.

Although there have been several studies comparing latent trait models for different testing applications, none have focused on a classroom examination system. Most model comparisons used simulated data or conveniently available standardized aptitude test data. The methodology described in the next chapter is designed to facilitate a practical comparison of latent trait models for test equating in a classroom examination system.

CHAPTER III

METHOD

SAMPLE

The test data is taken from an introductory communication course offered at Michigan State University. Most students taking the course are freshmen, some are sophomores and a few are juniors or seniors. The course is divided into sections of approximately fifty to seventy students taught by different instructors. Generally each instructor constructs his/her own mid-term examination. For research purposes the Fall 1979 mid-term was constructed by the course coordinator. All sections are given a common final examination. The final examination is constructed by the course coordinator from an item pool of approximately 600 items. Each final examination consists of 100 four-option multiple choice items. Test construction is based on principles derived from classical test theory. Items are preferred with high upper-lower discriminations and with difficulties of approximately .30. Each final examination reflects a balanced coverage of the ten content areas of the course. For a more detailed description of the course, the item pool, the test construction, as well as administrative and instructor training concerns; see Eisenberg and Book (1980).

The data for this study includes the Fall 1978, Winter 1979, Spring 1979 and Fall 1979 final examinations taken by 947, 820, 594 and 1082 examinees respectively. The Spring 1979 examination was chosen to

provide the base scale for calibrating abilities and item difficulties. Calibration required that the Spring 1979 examination share items with each of the other three examinations. Between the Spring 1979 examination and the Fall 1978, Winter 1979 and Fall 1979 examinations there were 49, 53 and 43 items in common, respectively.

In order to make the methodology and results of this study easier to assimilate, the overall study is subdivided into three parts. Study 1 corresponds to objectives i, ii, and iii in Chapter I. Study 1 investigates the equating consistency across models for the Fall 1978 and Winter 1979 final examinations and within and across models for the Fall 1979 final examination. For the purpose of investigating model consistency for various size examinee samples, independent random samples of 200, 600 and 800 examinees were selected from the 1082 examinees taking the Fall 1979 examination.

Study 2, corresponding to objective iv in Chapter I, attempts to provide a criterion for comparing models by equating the Fall 1979 final examination to itself using examinee samples differing in relative levels of ability. The Fall 1979 mid-term, rather than the final examination, was used to determine the ability groupings so that group assignment would be independent of errors of measurement on the final examination. This was the only use of the Fall 1979 mid-term examination.

The midterm examinations for Fall 1979 all had 37 items in common. The scores on the 37 item subtest were used to select the samples of students for the Fall 1979 final examination. Three separate sets of samples were selected. The first set was simply a random split of the 937 examinees for which both mid-term and final examination data were

available. The second set consisted of very low and very high ability groups. The groups were formed by placing examinees with scores below the median on the mid-term into the very low group and examinees with scores above the median into the very high group. Examinees at the median were randomly assigned to the two groups. The third set consisted of a low group and a high group designed to represent an ability split that was half of the difference in original standard deviation units between the very low and very high ability groups. The probabilities that examinees above and below the median should have of going into each of the groups were determined so that the expected difference in group means would be as desired. These probabilities were used with a list of random numbers to assign examinees to the low and high ability groups.

Study 3, corresponding to objective v in Chapter I, is designed to investigate the effect of different numbers of uncalibrated items upon ability estimates within examinee sections. The 43 items also on the Spring 1979 final were selected from the Fall 1979 final examination to simulate a mid-term length examination. Sixteen intact examinee sections ranging in size from 49 to 66 students provided the examinee sample totaling 917.

MODELS

The three models selected for comparison were the three-, two- and one-parameter or Rasch, logistic item response models. Each of these models specifies a particular mathematical relationship between person ability and the probability of a correct response to a given item (see Hambleton and Cook, 1977 for a good introduction to ICC models). The three-parameter model can be written:

$$P_g\ (\theta, a_g, b_g, c_g) = c_g + (1 - c_g)\ (1 + e^{-1.7\ a_g\ (\theta - b_g)})^{-1}, \quad (3.1)$$

where $P_g\ (\theta, a_g, b_g, c_g)$ is the probability of a correct response to item g by a person of ability θ,

c_g is the probability of examinees of extremely low ability giving a correct response to item g, i.e. the lower asymptote of the ICC,

b_g is the point on the ability scale where the slope of the ICC is maximum and is usually called the item's difficulty,

a_g is proportional to the slope of P_g at $\theta = b_g$ and is called the item's discrimination,

and e is the base of the natural logarithms and is approximately equal to 2.718.

The two-parameter model is identical to the three-parameter model when all of the lower asymptotes (c_g) are fixed at a constant, usually zero. The two-parameter model when $c_g = 0$ can be written:

$$P_g\ (\theta, a_g, b_g) = (1 + e^{-1.7\ a_g\ (\theta - b_g)})^{-1}, \quad (3.2)$$

where the notation is defined above.

If the model is further simplified by substituting $\bar{a}$, a common value for the item discriminations in a test, for a_g, the one-parameter model is obtained:

$$P_g\ (\theta, b_g) = (1 + e^{-1.7\ \bar{a}\ (\theta - b_g)})^{-1}. \quad (3.3)$$

A version of the one-parameter logistic model was developed independently by Rasch (1960). This version, called the Rasch model, is usually written with the constant 1.7 ã incorporated into the θ (b) scale and can be written:

$$P_g(\theta', b_g') = (1 + e^{-(\theta' - b_g')})^{-1}. \quad (3.4)$$

ESTIMATION OF MODEL PARAMETERS

The computer program LOGIST (Wood, Wingersky and Lord, 1976) was used to estimate ability and item parameters for the two- and three-parameter logistic models. Although LOGIST can be used to estimate parameters for the one-parameter model, BICAL, (Wright, Mead, and Bell, 1979) was used. One trivial difference between the programs is that the estimates for LOGIST are based on equations (3.1) - (3.3) while BICAL is based on equation (3.4). More importantly, BICAL is much easier to use than LOGIST and has many convenient and useful features not found in LOGIST.

SCALING MODEL PARAMETER ESTIMATES

The degree of uniqueness of the person ability and item parameter scales is determined by the item response models. One critical issue is the extent to which the parameters of a model are determined. For the three-parameter model, consider a single item, g, where $P_{g1}(\theta_1, a_1, b_1, c_1,)$ represents one scaling and $P_{g2}(\theta_2, a_2, b_2, c_2,)$ represents another scaling. We will consider what $P_{g1} = P_{g2}$ implies about the relationship between the two scales.

Notice immediately from equation (3.1) that the minima of the P_g's occur as their respective θ's approach negative infinity. The minimum

of P_{g1} is c_1, and the minimum of P_{g2} is c_2, but we are considering the same item so $c_1 = c_2$. In other words, the lower asymptotes of the item characteristic curves are completely determined.

Now consider the case where P_{g1} and P_{g2} equal a constant greater than c_g. For the probabilities to be equal, it can be seen from (3.1) and the fact that $c_1 = c_2$ that:

$$a_1 (\theta_1 - b_1) = a_2 (\theta_2 - b_2). \tag{3.5}$$

Rearranging terms algebraically we find that

$$\theta_1 - \frac{a_2}{a_1} \theta_2 = b_1 - \frac{a_2}{a_1} b_2. \tag{3.6}$$

But we are considering a single item with $P_{g1} = P_{g2}$ so a_1, a_2, b_1 and b_2 are all constants. Therefore there exists a constant K equal to each side of equation (3.6). We can write

$$\theta_1 = K + \frac{a_2}{a_1} \theta_2 \tag{3.7}$$

and

$$b_1 = K + \frac{a_2}{a_1} b_2. \tag{3.8}$$

Notice in the Rasch Model that (3.7) and (3.8) reduce to

$$\theta' = K + \theta_2' \tag{3.9}$$

and

$$b_1' = K + b_2', \tag{3.10}$$

since the a's are not modeled.

The LOGIST program solves the problem of the indeterminacy in the person ability and item difficulty scales by setting the average person ability to approximately zero and the standard deviation of ability to approximately one on any given computer run. Therefore, two separate LOGIST runs with different groups of examinees will result in scales that are linear transformations of one another. In this study we have different examinees but common items across tests so equation (3.8) is the appropriate one to examine, where now b_1 is on the scale defined on one test and b_2 is on the scale defined by another test.

Calibration requires estimation of the transformation represented by (3.8). This can be done in several ways. Perhaps the most obvious way is to use the item difficulty estimates to estimate the parameters of the regression equation.

$$b_1 = \alpha + \beta\, b_2, \tag{3.11}$$

where we have as many sets of $(\hat{b}_1, \hat{b}_2)$ as we have items in common between the two tests. Unfortunately, this is not a very good way to estimate the parameters of the transformation equation (Lord, personal communication, April 1980). The problem is that two inconsistent transformation equations will result depending on which test is regressed on the other. The magnitude of the discrepancy will depend on the correlation between the two sets of difficulty estimates. Identical equations result only if the correlation is one. The equations are increasingly different as the correlation departs from one.

One method (Warm, 1978, pp. 113-116) which avoids this problem is to let

$$\hat{\alpha} = \bar{b}_1 - (s_{b1} / s_{b2})\, \bar{b}_2 \tag{3.12}$$

and

$$\hat{\beta} = s_{b1} / s_{b2}, \tag{3.13}$$

where $\bar{b}_1$ is the mean of the $\hat{b}_1$'s,
$\bar{b}_2$ is the mean of the $\hat{b}_2$'s,
s_{b1} is the standard deviation of the $\hat{b}_1$'s,
and s_{b2} is the standard deviation of the $\hat{b}_2$'s.

Substituting $\hat{\alpha}$ and [illegible] for α and β in equation (3.11) provides a consistent transformation regardless of which test is chosen to define a base scale. Notice that the transformation dependent on (3.12) and (3.13) can be thought of as a regression equation where the correlation between the b's has been replaced by one. Once $\hat{\beta}$ has been determined, the item discriminations are transformed by the equation obtained by substituting $\hat{\beta}$ for β in the equation

$$a_1 = a_2/\beta \tag{3.14}$$

An alternative method of estimating α and β, developed by the author, uses both the item discriminations and difficulties. This method was obtained by comparing equations (3.8) and (3.11), and noticing that they are identical if $K = \alpha$ and

$$\frac{a_2}{a_1} = \beta. \tag{3.15}$$

Then a reasonable estimate of β is the average of the ratios $\hat{a}_2/\hat{a}_1$. Here there are as many $(\hat{a}_1, \hat{a}_2)$ as there are items in common between the two tests. An estimate of α can be obtained by substituting $\hat{\beta}$ for β, the mean of the b_1's for b_1 and the mean of the b_2's for b_2 in equation (3.11)

and solving for α. Each method will result in a generally different linear equation for converting abilities and difficulties from the second scale to the first. Other methods could be developed that also use both the difficulties and discriminations simultaneously. However, the best method of combining the information contained in the discriminations and difficulties is not obvious. However, Lord (personal communication, April 1980) is currently working on robust estimators for the transformation equation.

The BICAL computer program solves the indeterminacy of the Rasch scale by setting the mean of the items in any given test to zero. Substituting the mean of the b_1''s for b_1' and the mean of the b_2''s for b_2' and solving for K in equation (3.10) results in

$$K = \overline{b_1'} - \overline{b_2'}. \qquad (3.16)$$

This estimated value of K can be simply added to each b_2' to place all the items on a common scale.

EQUATING

Once scaling has been accomplished, test equating is straightforward. If ability scores are being compared, then the same transformation that was applied to item difficulties can be applied to abilities to form ability measures on the same scale.

If raw-score equating is desired then a further complication is introduced; it is necessary to set up a correspondence between raw scores on different tests through the ability estimates. One method of equating uses true-score equating applied to raw scores. The true score on a test for a given ability is simply the expected value of the observed score at

that ability level. This is simply the sum of the item characteristic curves at that ability, i.e.,

$$T(\theta) = E(X|\theta) = \Sigma P_g(\theta), \qquad (3.17)$$

where the summation is over all the items in a test. Equating is simply a matter of tabling or graphing the relation of $\Sigma \hat{P}_g(\theta)$ on one test to that of $\Sigma \hat{P}_g(\theta)$ on a second test for many fixed values of θ, where $\hat{P}_g(\theta) = P_g(\theta, \hat{a}_g, \hat{b}_g, \hat{c}_g,)$ (Lord, 1977).

PROCEDURE

Separate estimates for each of the samples mentioned above were obtained for the Rasch model using BICAL. For Study 1 and 2, examinees with less than 40 of 100 items correct or with person t fits above 2.00 (see Wright, Mead, and Bell, 1979, p. 15) were deleted from the item calibration sample. From 1/2% to 5% of the examinees in a given sample were removed under these criteria. It was felt that persons with scores of less than 40 were likely to be guessing on many items and thereby adding error to item calibrations. Persons with large t fit values were removed since they were demonstrating unusual response patterns which could adversely affect item parameter estimation. For the same reasons in Study 3, examinees with less than 17 of the 43 items correct or person t fits above 2.00 were removed from the calibration sample.

In Study 1 all item difficulties were transformed to the scale defined by the Spring 1979 examination using equations (3.10) and (3.16). The person abilities were then transformed to a standard scale designed so that the mean ability of the Spring 1979 examinees is 500 and the standard deviation is 100. The transformation equation is

$$S = \frac{\hat{\theta} - 1.04}{.60} (100) + 500 \qquad (3.18)$$

where S is the scaled score,

$\hat{\theta}$ is the estimated ability on the Spring 1979 scale,

1.04 is the person ability mean on the Spring 1979 scale,

and .60 is the ability standard deviation on the Spring 1979 scale.

Separate estimates for each sample in Study 1 and 2 were obtained for the two-parameter model using LOGIST. Person abilities and item difficulties and discriminations for all samples in Study 1 were transformed to the scale determined by the Spring 1979 examination. The method using item difficulties and the average ratio method using item discriminations and average difficulties were used to estimate α and β in equation (3.11). This resulted in two different transformations for the two-parameter model. The additional transformation

$$S = (100)\,\theta + 500$$

was applied to the scaled estimates to facilitate comparison between the Rasch and two-parameter models.

LOGIST was used to obtain separate estimates for the three-parameter model on the Fall 1978, Winter 1979, Spring 1979 and Fall 1979 final examination total samples. No calibrations were done.

For Study 2, the three pairs of ability samples taken from the Fall 1979 final examination were used to equate the Fall 1979 examination to itself. For each of the three pairs, one sample in each pair was calibrated to the other using only those 43 items that were used to calibrate

the Fall 1979 examination to the Spring 1979 scale. The calibrated abilities were used for true-score to true-score equating. The two methods of item calibration for the two-parameter model provided the basis for two methods of equating and the Rasch model provided a third.

For Study 3, BICAL was used to obtain Rasch item difficulties and ability estimates using the 43 items selected from the Fall 1979 final and the entire group of Fall 1979 examinees for which section membership was known. Separate BICAL computer runs were also made for each of the 16 sections for the 43 item subtest. Average item difficulties were calculated for subsets of 7, 12, 17, 22, 27, 32 and 37 items from the 43 total items. The averages were calculated for the total group and for each section. This allowed comparisons of calibration constants and hence ability estimates, acting as though there were different numbers of calibrated items within the total group of 43 items.

CHAPTER IV

RESULTS

The results will be presented in four sections with separate discussions for each study. The first section contains descriptive data for the various tests and examinee samples. The next three sections present the results of parameter estimation, item calibration and test equating.

DESCRIPTIVE DATA

Table 4.1 contains the basic descriptive information for Study 1. For each final examination and examinee sample, the number of examinees and test mean, standard deviation and KR-20 reliability are reported. In Table 4.1 all test statistics are based on simple number-right scoring. The means for the three random samples of the Fall 1979 examination range from 68.48 to 69.01 and the standard deviations range from 13.14 to 13.56.

Table 4.2 contains the same descriptive information for the Study 2 data. The means range from 61.66 for the very low ability sample to 76.65 for the very high ability sample. The standard deviation of the very high ability sample is smaller than the others due to a ceiling effect on the final examination. Based on the total Fall 1979 examinee sample standard deviation of 13.37, the two random samples differ by

TABLE 4.1

DESCRIPTIVE DATA FOR STUDY 1

Term	Sample	# Examinees	Raw Mean	Raw S.D.	KR-20
Fall 1978	Total	947	69.26	12.95	.89
Winter 1979	Total	820	68.83	12.05	.87
Spring 1979	Total	594	70.62	11.62	.87
Fall 1979	Total	1082	68.67	13.37	.90
Fall 1979	Random	200	69.01	13.14	.89
Fall 1979	Random	600	68.56	13.56	.90
Fall 1979	Random	800	68.48	13.30	.90

TABLE 4.2

DESCRIPTIVE DATA FOR STUDY 2

Term	Sample	# Examinees	Raw Mean	Raw S.D.	KR-20
Fall 1979	Random	481	69.49	13.19	.90
Fall 1979	Random	456	68.72	12.82	.89
Fall 1979	Very Low	471	61.66	11.84	.85
Fall 1979	Low	461	64.76	12.77	.88
Fall 1979	High	476	73.34	11.81	.88
Fall 1979	Very High	466	76.65	9.28	.82

only .06 standard deviation units. The low and high examinee ability samples differ by .64 units and the very low and very high samples differ by 1.12 units.

The descriptive data for Study 3 is presented in Table 4.3. The average Rasch ability ($\hat{\theta}$) for the entire sample is 1.11 with a standard deviation of .67. The BICAL program automatically sets the average difficulty of the items in any given test to zero. It can be seen that the average person ability is considerably higher than the average item difficulty. The examinee section means range from .87 to 1.44. Standard deviations range from .52 to .85. Based on the overall standard deviation of .67, the two most extreme sections differ by an average ability of .85 standard deviation units.

PARAMETER ESTIMATION

The three-parameter model has unacceptable lower asymptote estimates (c_g) for the data used in this study. Lord (1975b) has shown that c_g is poorly estimated when $(b_g - 2/a_g)$ is less than negative two. Consequently when $(b_g - 2/a_g) \leq -2$, LOGIST fixes c_g at the average of the other unfixed c's. For the Fall 1978, Winter 1979, Spring 1979 and Fall 1979 examinations 98, 98, 98, and 99 of 100 c's were fixed at the average of the c's for the remaining 2, 2, 2, and 1 items. The averages were .250, .205, .130 and .075, respectively. In effect the three-parameter model was reduced to an inconsistent two-parameter model. Therefore, the three-parameter model was given no further consideration.

On the whole, estimation of parameters for the two-parameter model was acceptable. Since discriminations (a_g) can approach zero or infinity, LOGIST places the restriction $.01 \leq a_g \leq 2.0$. No more than one of

TABLE 4.3

DESCRIPTIVE DATA FOR STUDY 3

Section Number	Mean $\hat{\theta}$	S.D. $\hat{\theta}$	# Examinees
1	1.02	.65	60
2	1.25	.54	54
3	1.10	.67	65
4	1.27	.60	49
5	1.07	.63	60
6	.87	.56	57
7	1.23	.61	57
8	.98	.76	57
9	1.16	.75	56
10	1.40	.85	55
11	.95	.66	54
12	1.29	.81	55
13	1.11	.68	57
14	1.16	.52	53
15	1.44	.72	66
16	1.14	.68	62
All	1.11	.67	917

100 discriminations was fixed at .01 and no items approached 2.0 in any two-parameter LOGIST run. There were at least a few implausible item difficulty estimates in every examination.

Tables 4.4 and 4.5 summarize the two-parameter model item difficulty ($\hat{b}$) estimates for Studies 1 and 2. Items with difficulties between -3 and +2 should be reasonably well estimated. Recall that LOGIST sets the ability and item difficulty scale so that the mean person ability is approximately zero and the person ability standard deviation is approximately one. Items with difficulties much below -3 or above +2 are not likely to be very well estimated, since there is relatively little useful item response data at those extremes. Therefore, the number of items with difficulties between -3 and +2 is an indication of the number of items that are likely to have reasonably well estimated parameters. It can be seen from Table 4.5 that for the very high ability sample only 64 of 100 items were in that range. For the remaining samples in Tables 4.4 and 4.5, between 84 and 96 of 100 item difficulties were in the range of -3 to +2.

Tables 4.4 and 4.5 also contain the difficulty estimates for the lowest and highest of the 100 items in each test sample. Notice that item parameter estimates obtained from LOGIST can obtain absurdly low or high values. Such items are essentially useless for measurement in the present situation. Many have extremely flat ICC slopes leading to arbitrary and unstable difficulty estimates.

Rasch model estimates obtained from BICAL were all plausible. However, the item tests of fit provided by BICAL are highly dependent on sample size. For example, for the examinee samples of 200, 600, 800 and 1082 from the Fall 1979 examination, 8, 13, 25, and 29 items,

TABLE 4.4

TWO-PARAMETER ITEM DIFFICULTY ($\hat{b}$) STATISTICS FOR STUDY 1

Term	Sample	# Examinees	# Items with $\hat{b}$'s in Range -3 to +2	Low $\hat{b}$	High $\hat{b}$
Fall 1978	Total	947	89	-13.12	31.16
Winter 1979	Total	820	84	-34.97	2.48
Spring 1979	Total	594	88	-94.54	3.36
Fall 1979	Total	1082	94	- 6.05	1.46
Fall 1979	Random	200	92	-19.25	1.31
Fall 1979	Random	600	96	- 5.57	2.61
Fall 1979	Random	800	92	-11.36	2.20

TABLE 4.5

TWO-PARAMETER ITEM DIFFICULTY ($\hat{b}$) STATISTICS FOR STUDY 2

Term	Sample	# Examinees	# Items with $\hat{b}$'s in Range -3 to +2	Low $\hat{b}$	High $\hat{b}$
Fall 1979	Random	481	88	- 5.67	1.86
Fall 1979	Random	456	90	- 11.75	1.47
Fall 1979	Very Low	471	90	- 5.49	4.64
Fall 1979	Low	461	88	- 9.95	2.84
Fall 1979	High	476	87	- 92.52	1.15
Fall 1979	Very High	466	64	-120.21	.93

respectively, were identified as having an overall t fit greater than 2.0. It is clear that the fit statistics are more appropriate for item selection than for testing model fit unless some means of taking sample size into account is used.

In conclusion, estimates were successfully obtained for the Rasch and two-parameter models. The next two sections on calibration and equating provide information on the consistency, accuracy and usefulness of the estimates.

CALIBRATION

Rasch Model

For the Rasch model, the Study 1 samples were separately calibrated to the Spring 1979 scale by the method described in Chapter III. Table 4.6 presents the calibrating constants, residual and standardized residual statistics and the numbers of items in the calibrating samples.

The residual is simply the difference for the common items between the actual Spring 1979 item difficulties and the item difficulties calibrated from another test scale. The residual mean is necessarily zero. A totally error-free calibration would result in a residual standard deviation of zero.

The standardized residual is simply a residual divided by the standard error of that residual. The standard error of a residual is

$$S_R = (S_1^2 + S_2^2)^{1/2} \qquad (4.1)$$

where S_1^2 is the error variance for an item in test 1 and S_2^2 is the error variance for the same item in test 2. In general, different items will have different residual standard errors.

TABLE 4.6

RASCH MODEL CALIBRATION FOR STUDY 1

Term	Sample	# Examinees	K	Residual S.D.	Standardized Residual Mean	Standardized Residual S.D.	# Calibrating Items
Fall 1978	Total	947	.072	.39	.06	3.01	49
Winter 1979	Total	820	.131	.31	.02	2.30	53
Fall 1979	Total	1082	.172	.37	-.05	2.93	43
Fall 1979	Random	200	.179	.37	-.08	1.83	43
Fall 1979	Random	600	.170	.37	-.08	2.57	43
Fall 1979	Random	800	.172	.36	-.05	2.70	43

The form of the equation which transforms item difficulty estimates from one scale ($\hat{b}$) to the Spring 1979 scale ($\hat{b}_s$) is:

$$\hat{b}_s = \hat{b} + K.$$

The mean of the standardized residuals should be approximately zero and the standard deviation should be about 1.0. It can be seen in Table 4.6 that the means are all close to zero, with all discrepancies less than .10. However, the standard deviations are up to three times larger than expected. This implies that there is more error in the calibrating links than can be explained by the model.

In Study 2, the random halves were put on a common scale, as were the low and high sample and the very low and very high sample. It was not necessary to place the three sets of samples on a scale common to all, since true-score equatings were made only between pairs of samples in a set. Therefore one random sample was arbitrarily placed on the same scale as the other, the low ability sample was calibrated to the high and the very low ability sample was calibrated to the very high ability sample.

The Rasch model calibration statistics for Study 2 are presented in Table 4.7. Although the tests are being calibrated to one another as though only 43 items are common to both, actually they are identical 100 item tests. Therefore, the scales should be the same from the beginning, since BICAL standardizes on item difficulties rather than on examinee abilities. Then obviously the calibration constant, K, should be zero in each case. In fact, the constants are small, ranging from -.061 to +.001. Notice that despite the fact that the very low to very high ability sample calibrating constant is the closest to zero (.001), the residual standard deviation, the standardized residual mean and the standardized residual standard deviation are the largest. This means that although the differences in item difficulties across samples averaged out very close to the proper value, there was less

TABLE 4.7

RASCH MODEL CALIBRATION FOR STUDY 2

Term	Sample	# Examinees	K	Residual S.D.	Standardized Residual Mean	Standardized Residual S.D.	# Calibrating Items
Fall 1979	Random	481	-.061	.17	-.01	1.02	43
Fall 1979	Low	461	-.028	.25	.08	1.51	43
Fall 1979	Very Low	471	.001	.33	.17	1.87	43

The form of the equation which transforms item difficulty estimates from one scale ($\hat{b}$) to the other scale ($\hat{b}_s$) is:

$$\hat{b}_s = \hat{b} + K.$$

consistency compared to the other samples. The effects of calibration and item difficulty estimate error due to different examinee ability samples will be seen in the test equating section.

Two-Parameter Model

For the two-parameter model, both methods of calibration previously described were separately applied to the various samples in Studies 1 and 2. As with the Rasch model, the Spring 1979 examination provided the common scale for the Study 1 samples. Table 4.8 contains the calibrating constants estimated from the item difficulties method (B Method) and from the ratio of item discriminations method (A Method) for Study 1. It also contains correlations between the difficulty estimates and numbers of common items between Spring 1979 and each of the other examinations used for calibration. A comparison of Table 4.6 and 4.8 reveals that three items in the Fall 1978 examination and two items in the Winter 1979 examination were not used for calibration. Scatterplots of item difficulties estimated for the common items revealed that all five items were clearly outliers. All five greatly affected the calibration equations, so they were removed and the calibration equations were again calibrated. Figures A1 to A6 in the Appendix are scatterplots of the difficulty estimates for the items actually used for calibration in Study 1. Table A1 in the Appendix contains the difficulty estimates for the items that were removed.

The potential effect of outliers on item calibration can be seen in Table 4.8. The random sample of 200 resulted in one outlying item (see Table A1). Notice that the calibrations are quite different with just the one outlier removed. Also notice that the correlation between

TABLE 4.8

TWO-PARAMETER MODEL CALIBRATION FOR STUDY 1

Term	Sample	# Examinees	B MTH $\hat{\alpha}$	B MTH $\hat{\beta}$	A MTH $\hat{\alpha}$	A MTH $\hat{\beta}$	Correlation $\hat{b}$'s	# Calibrating Items
Fall 1978	Total	947	.668	1.536	.107	1.157	.86	46
Winter 1979	Total	820	.006	1.079	.018	1.089	.87	51
Fall 1979	Total	1082	-.043	.995	.178	1.138	.57	43
Fall 1979	Random	200	-.535	.642	.522	1.295	.46	43
Fall 1979	Random	200	-.125	.968	.393	1.317	.58	42
Fall 1979	Random	600	-.041	1.004	.154	1.132	.58	43
Fall 1979	Random	800	-.083	.949	.178	1.115	.51	43

The equations to transform item parameter estimates from one scale ($\hat{a}$, $\hat{b}$) to the Spring 1979 scale ($\hat{a}_s$, $\hat{b}_s$) are:

$$\hat{b}_s = \hat{\alpha} + \hat{\beta}\hat{b}$$

and

$$\hat{a}_s = \hat{a} / \hat{\beta}$$

the item difficulties changes from .46 to .58 when the one item is removed. It can be seen from Table 4.8 that there is some discrepancy between the calibration equations based on difficulties and those based on average discrimination ratios for all Fall 1979 samples. The Fall 1979 sample of 200 calibrated using all 43 items results in equations that are quite different from those based on the larger samples, particularly for the difficulty method. The effects of these discrepancies on equating will be seen in the next section.

Table 4.9 contains the two-parameter item calibration information for the Study 2 examinee samples. Five items were removed before calibration of the very low ability sample to the very high ability sample. The difficulties of the five items were more than five standard deviations below the person ability mean and were clearly outliers on the scatterplot of item difficulties for the very low and very high ability samples. The correlation between the item difficulty estimates changed from .52 to .84 when the five items were removed. The scatterplots of the item difficulties for the items used for calibration in Study 2 are Figures A7 to A9 in the Appendix.

TEST EQUATING AND ABILITY SCALING

Study 1

The Fall 1978, Winter 1979 and Fall 1979 examinations were equated to a standard reference scale. The standard scale was defined to have a mean of 500 and standard deviation of 100 for the Spring 1979 examinee sample. As described in Chapter III, linear transformations of the calibrated Rasch and two-parameter model abilities were used to transform the abilities to the reference scale.

TABLE 4.9

TWO-PARAMETER MODEL CALIBRATION FOR STUDY 2

Term	Sample	# Examinees	B MTH $\hat{\alpha}$	B MTH $\hat{\beta}$	A MTH $\hat{\alpha}$	A MTH $\hat{\beta}$	Correlation $\hat{b}$'s	# Calibrating Items
Fall 1979	Random	481	- .074	.967	.084	1.062	.81	43
Fall 1979	Low	461	- .641	1.022	- .598	1.053	.79	43
Fall 1979	Very Low	471	-1.310	1.100	-1.412	1.013	.84	38

The equations to transform item parameter estimates from one scale ($\hat{a}$, $\hat{b}$) to the other scale ($\hat{a}_s$, $\hat{b}_s$) are:

$$\hat{b}_s = \hat{\alpha} + \hat{\beta}\,\hat{b}$$

and

$$\hat{a}_s = \hat{a} / \hat{\beta}.$$

In the figures that follow, axes labeled "scaled scores" correspond to the standard reference scale and the axes labeled "expected raw scores" correspond to true scores. Figure 4.1 provides an indication of the degree of correspondence between the reference scales determined by the Rasch and two-parameter models. It can be seen that the scaling has resulted in reasonably equivalent scales, although the transformation based on the Rasch model generally assigns a higher expected raw score than the transformation based on the two-parameter model to the same reference scale score.

Figures 4.2 through 4.7 compare the equatings based on the Rasch and two scalings of the two-parameter model for the Fall 1978, Winter 1979, Fall 1979 total samples and random 200, 600 and 800 examinee samples of the Fall 1979 examination, respectively. The equatings based on the sample of 200 examinees use the calibration obtained with the one outlier removed. The variable labeled BMTH in the figures is the equating influenced by the calibration method which uses the item difficulty values. The variable labeled AMTH reflects the alternative calibration method which estimates the calibration weights using the average ratio of the item discrimination values and the average item difficulties. In Figures 4.2 through 4.7 the solid line represents the Rasch equating, the plus marks represent the equating using the two-parameter model calibrated by the difficulty method, and the circles represent the two-parameter equating with calibration by the alternate method.

The Fall 1978 equating in Figure 4.2 shows the same relationship between the Rasch and two-parameter equatings evidenced by the Spring 1979 equating. The Rasch equating produces higher true scores at most ability levels. The two calibrations of the two-parameter model produced equatings that are somewhat different from one another.

FIGURE 4.1: SPRING 1979 EQUATING

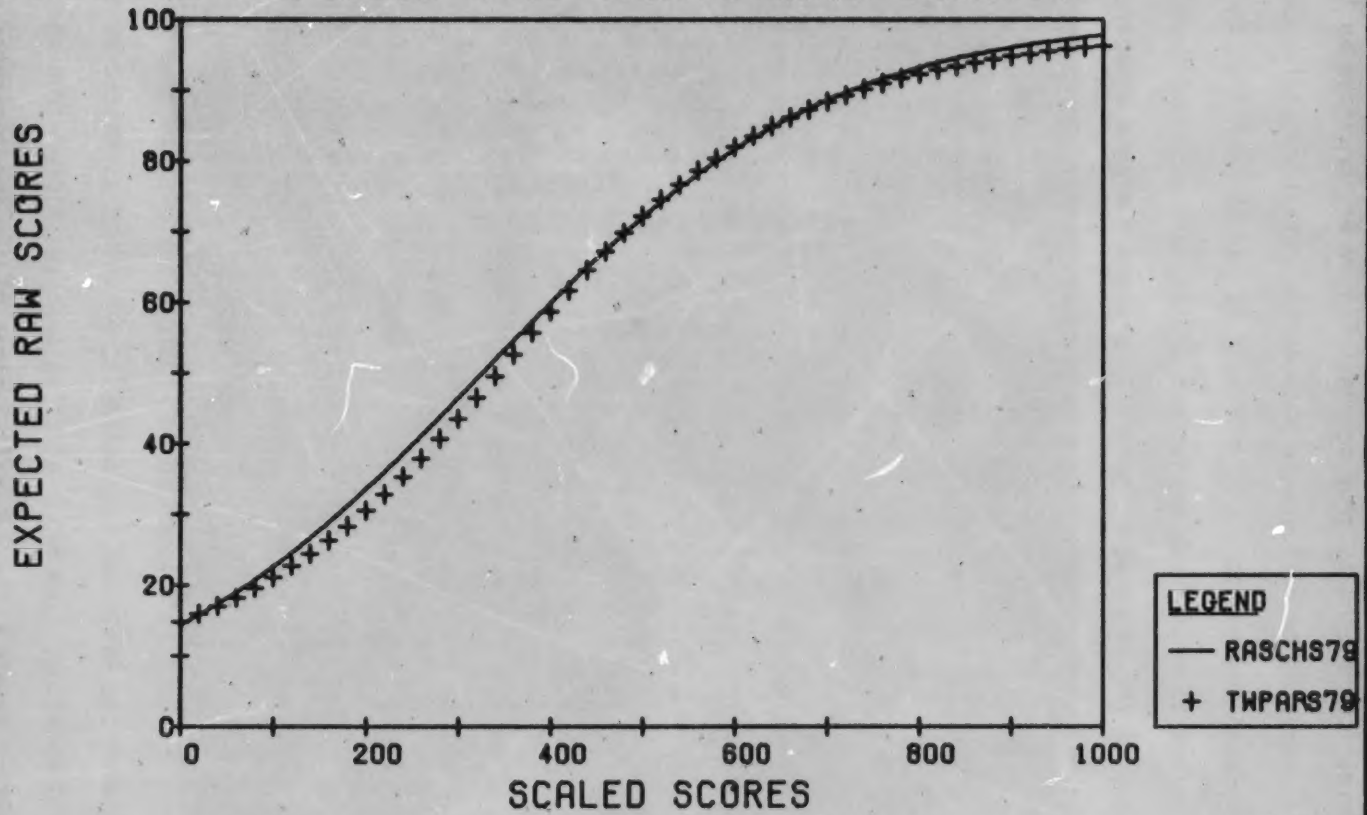


FIGURE 4.2: FALL 1978 EQUATING

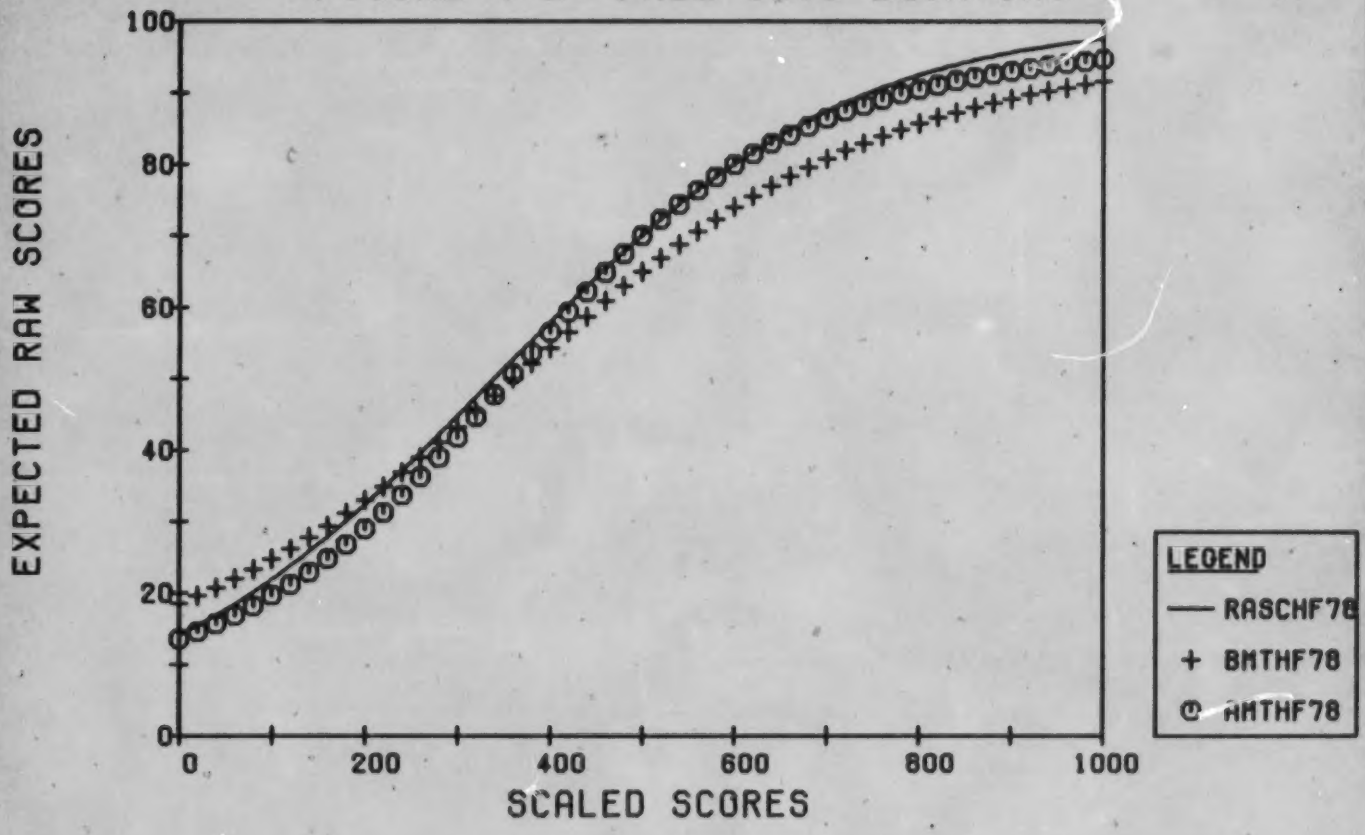


FIGURE 4.3: WINTER 1979 EQUATING

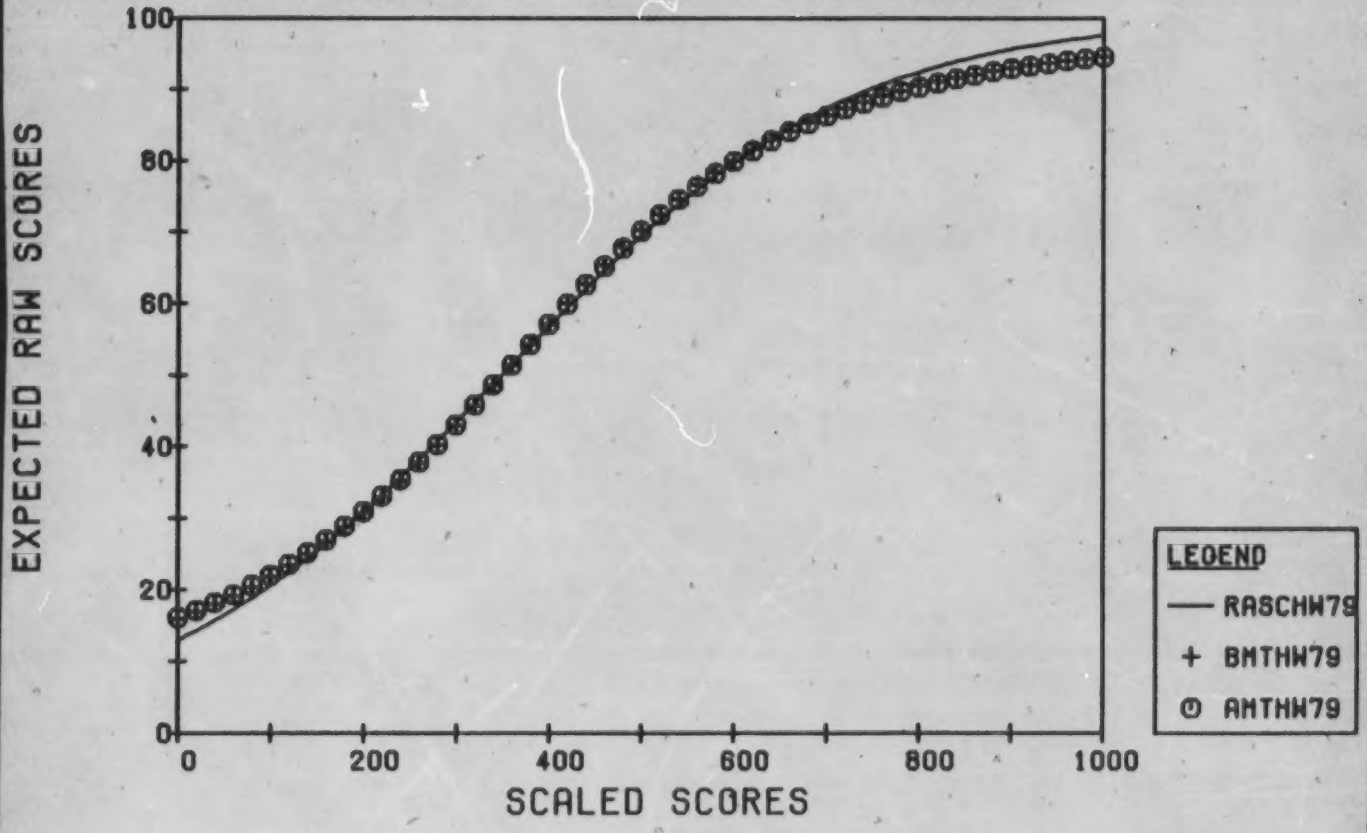


FIGURE 4.4: FALL 1979 EQUATING

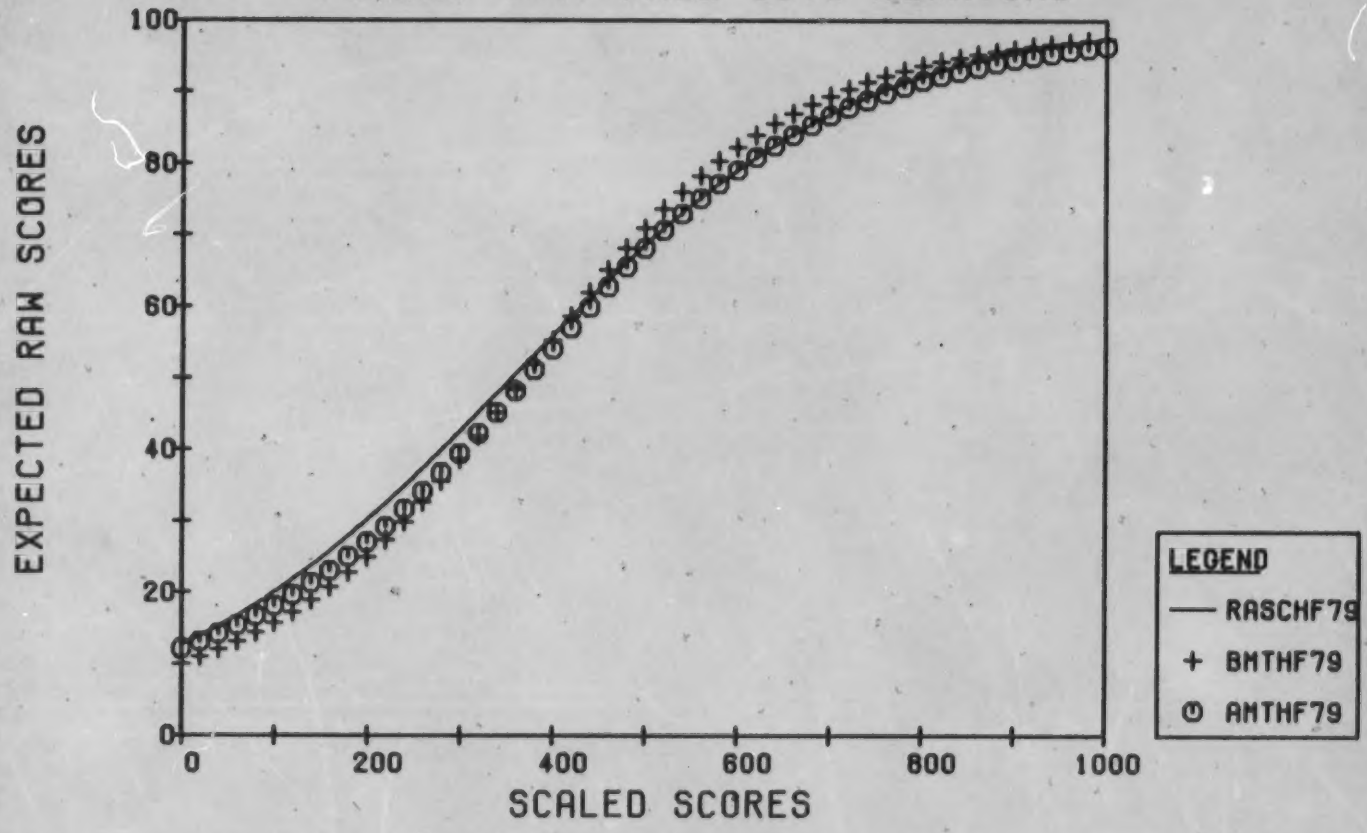


FIGURE 4.5: FALL 1979 SAMPLE OF 200 EQUATING

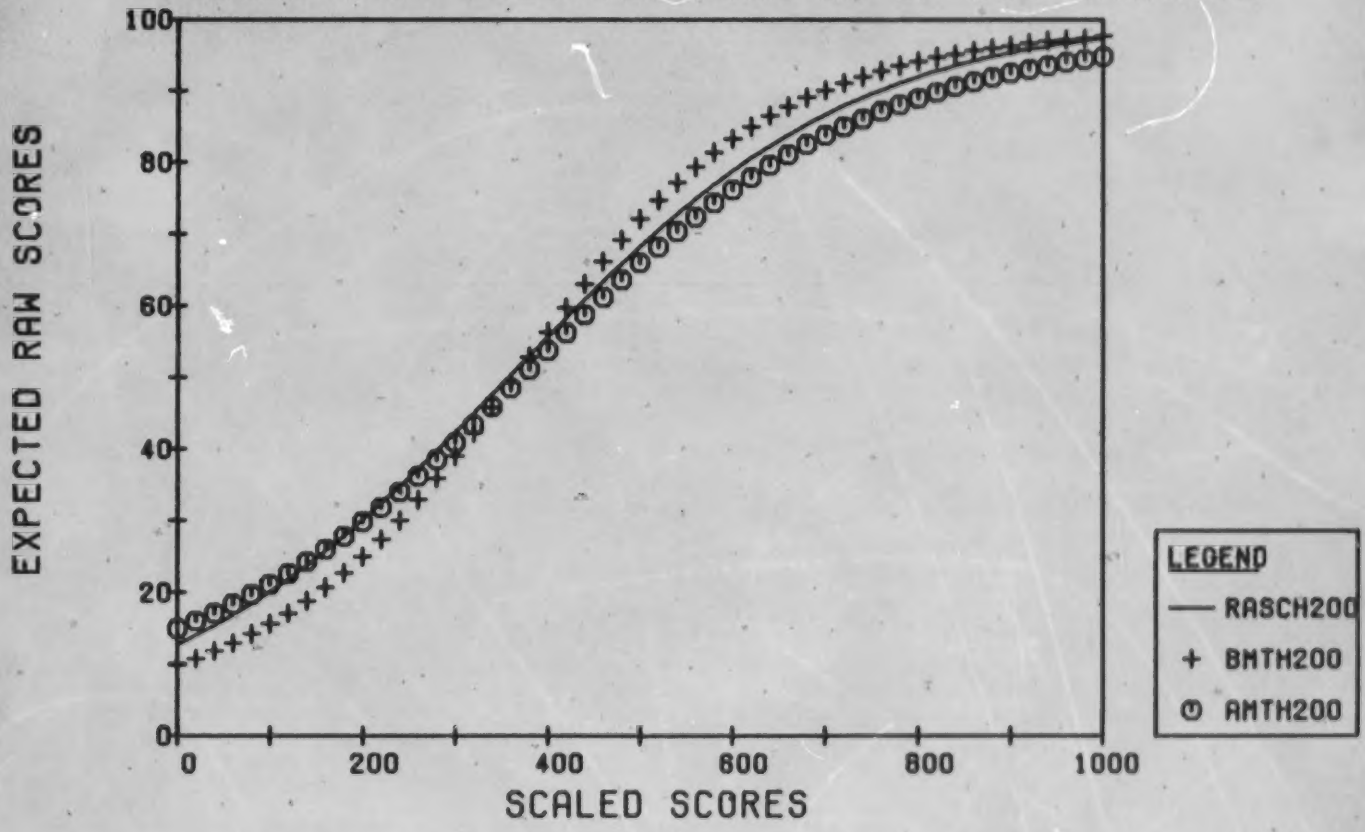


FIGURE 4.6: FALL 1979 SAMPLE OF 600 EQUATING

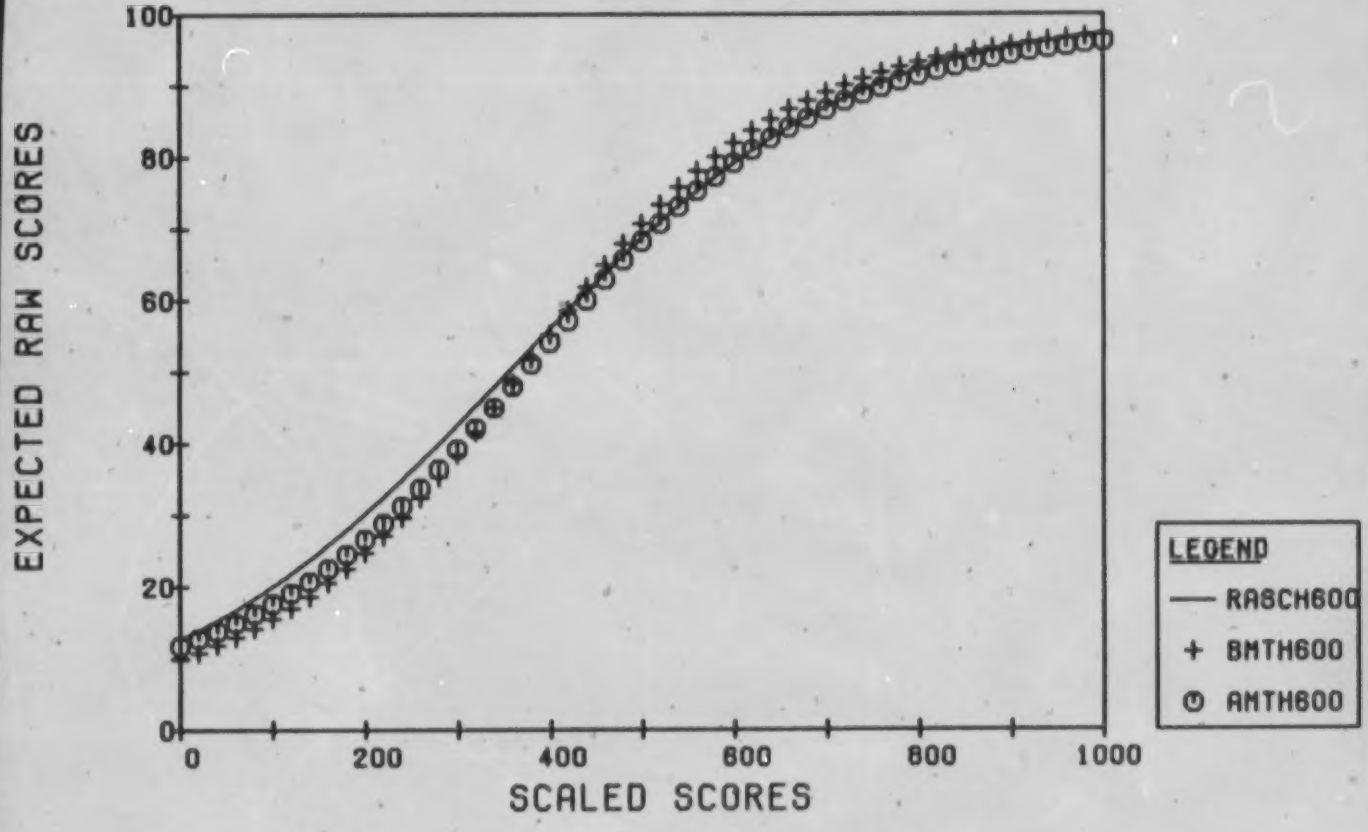
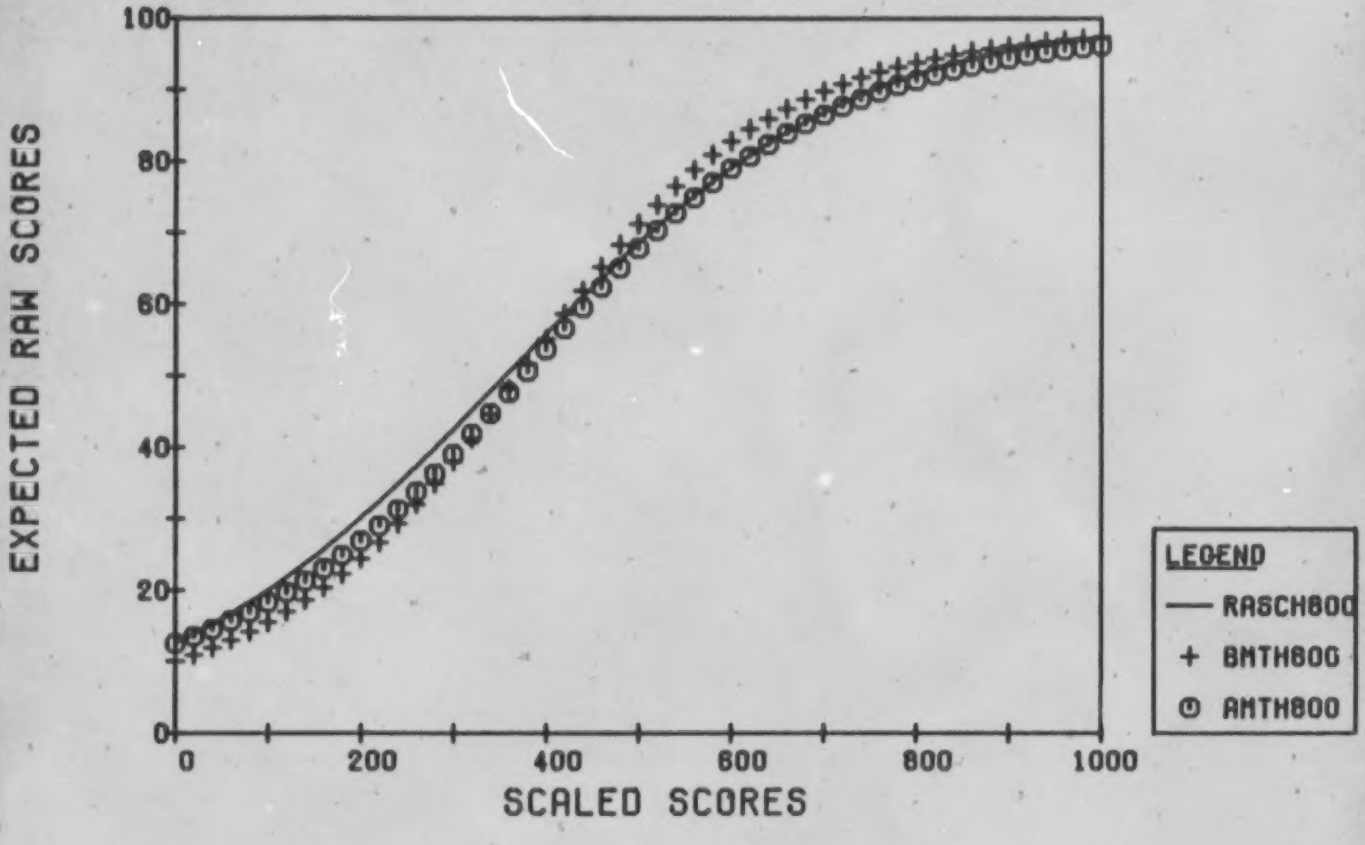


FIGURE 4.7: FALL 1979 SAMPLE OF 800 EQUATING



In Figure 4.3 the Winter 1979 equatings show a different pattern. The two-parameter equatings are very similar to one another. However, the Rasch equating produces lower true scores for low abilities and higher true scores for high abilities compared to the two-parameter equatings.

The Fall 1979 equatings based on the total examinee sample in Figure 4.4 show some divergence between the three methods. The Rasch model tends to give higher true-score equatings of lower abilities. The two methods of equating based on the two-parameter model diverge more than in the previous figure.

Equatings of the Fall 1979 examination to the reference scale based on random examinee samples of 200, 600 and 800 are presented in Figures 4.5, 4.6 and 4.7, respectively. Notice in Figure 4.5 that the two-parameter equating based on the item difficulty method of calibration shows the greatest divergence from the other equatings.

Figures 4.8, 4.9 and 4.10 display the consistency of equating within models but across examinee samples from the Fall 1979 examination. The Rasch model equatings in Figure 4.8 show remarkable consistency across the examinee samples. A small divergency of the two-parameter equating based on the item difficulty method of item calibration for the examinee sample of 200 can be seen in Figure 4.9. In Figure 4.10 the equatings, based on the item discrimination ratio method of calibration of the two-parameter model, show more divergence than the Rasch model equatings and more divergence than the equatings based on the other two-parameter method for the examinee sample of 200.

Although the equating figures give a useful overview of the test equatings, quantitative comparisons can convey information with more

FIGURE 4.8: RASCH EQUATING FALL 1979 SAMPLES

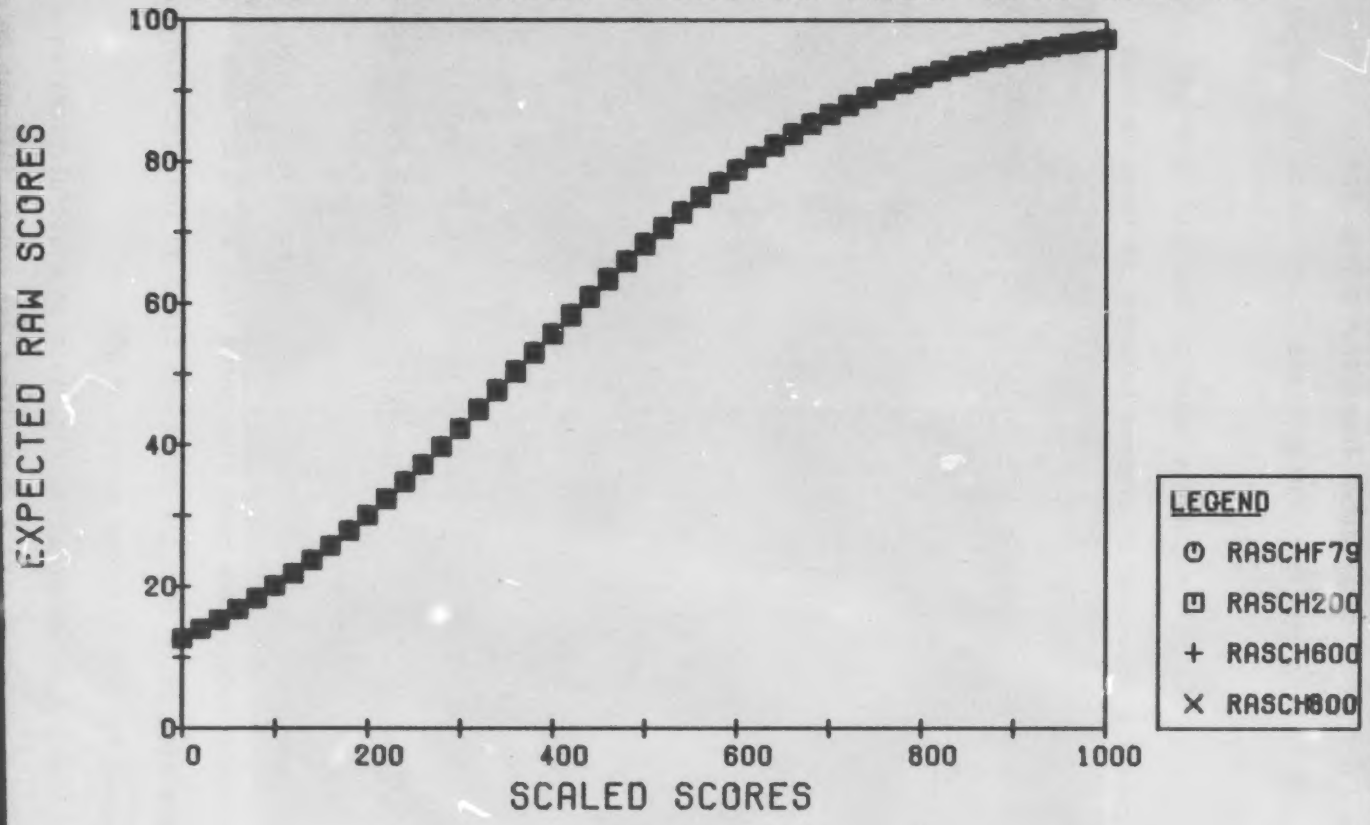


FIGURE 4.9: B METHOD EQUATING FALL 1979 SAMPLES

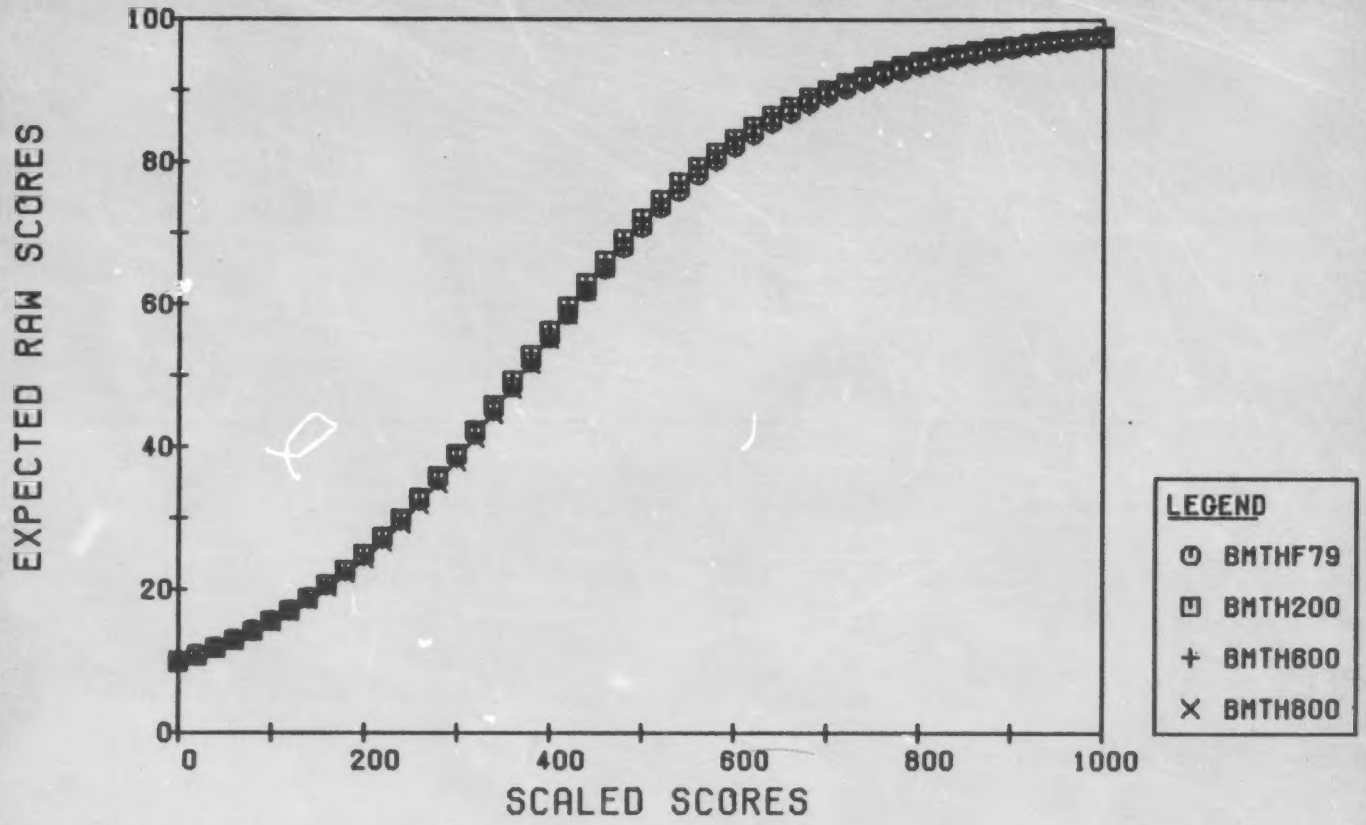
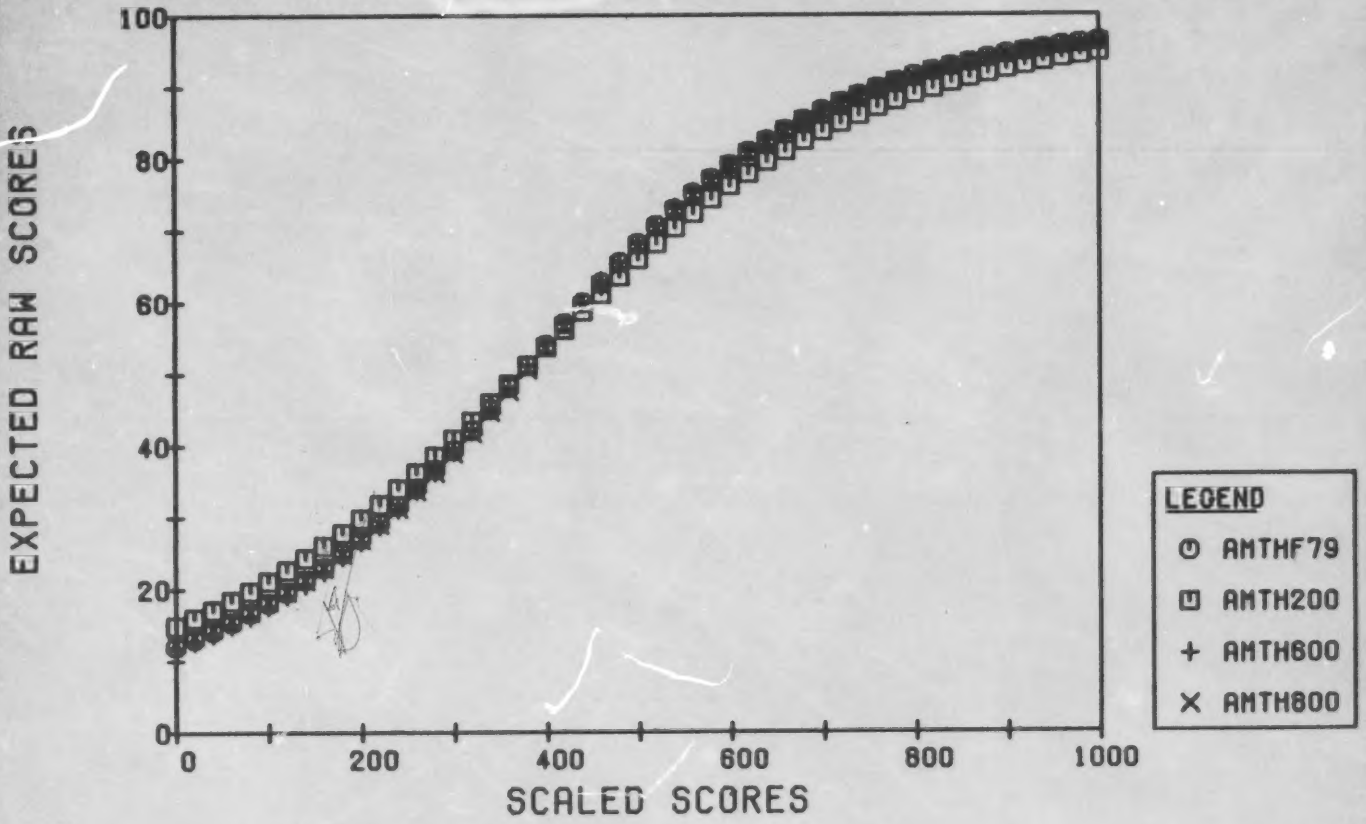


FIGURE 4.10: A METHOD EQUATING FALL 1979 SAMPLES



precision. Tables 4.10 and 4.11 provide quantitative comparisons of the three equating methods. Table 4.10 corresponds to Figures 4.2 through 4.7 and Table 4.11 corresponds to Figures 4.8, 4.9 and 4.10, The bias, standard deviation of difference and total inconsistency (or error) are related as follows:

$$\sum_j d_j^2/k = \sum_j (d_j - \bar{d})^2/k + \bar{d}^2 \qquad (4.2)$$

(Total Error) = (Variance of Difference) + (Squared Bias)

where d_j is the difference between true-score equatings at ability level j, $\bar{d} = \sum_j d_j/k$, and the summation is over the k ability levels uniformly distributed across the range investigated. For the samples represented in Tables 4.10 and 4.11, the ability ranged from zero to 1000 from which 201 points were taken in increments of five. This corresponds to a uniform sampling of ability between -5 and +5 standard deviations of mean ability based on the Spring 1979 examination. Notice that equation (4.2) differs from equation (2.1) in that differences are weighted uniformly across the ability scale in (4.2). Also using equation (4.2) results in estimates of the bias and the standard deviation of difference on the test true-score scale.

Table 4.10 quantifies the inconsistencies in equating across models that were noted in the discussion of the figures. For comparison it is interesting to note that the difference between the Rasch and two-parameter equating for the Spring 1979 examination (Figure 4.1) is reflected in a bias of 1.08, a standard deviation of difference of 1.13 and a total inconsistency of 2.44. The mean square difference reflects both the fact that the average Rasch model equating is 1.08 true-score points higher than the two-parameter equating and the fact that the standard deviation of the differences in equating is 1.13.

TABLE 4.10

EQUATING INCONSISTENCY ACROSS MODELS

				Rasch - B Method			Rasch - A Method			B Method - A Method		
Term	Sample	# of Examinees	# Calibrating Items	Bias	S.D Diff.	Total	Bias	S.D. Diff.	Total	Bias	S.D. Diff.	Total
Fall 1978	Total	947	46	3.32	3.66	24.45	1.73	1.11	4.21	-1.60	4.14	19.65
Winter 1979	Total	820	51	.32	1.48	2.30	.34	1.55	2.50	.02	.14	.02
Fall 1979	Total	1082	43	.50	2.95	8.98	1.28	1.06	2.77	.78	2.05	4.81
Fall 1979	Random	200	43	-1.10	8.52	73.74	2.68	1.75	10.24	3.78	9.56	105.75
Fall 1979	Random	200	42	- .15	3.40	11.58	1.60	1.51	4.84	1.75	4.70	25.17
Fall 1979	Random	600	43	.78	2.91	9.07	1.46	1.21	3.60	.68	1.81	3.75
Fall 1979	Random	800	43	.48	3.29	11.03	1.43	1.13	3.32	.94	2.42	6.75

Total inconsistency is equal to the square of the bias plus the square of the standard deviation of the difference.

TABLE 4.11

EQUATING INCONSISTENCY WITHIN MODELS

Comparison Based on Samples of Size	RASCH			TWO-PARAMETER B METHOD			TWO-PARAMETER A METHOD		
	Bias	S.D. Diff.	Total	Bias	S.D. Diff.	Total	Bias	S.D. Diff.	Total
200 and 1082	.146	.132	.039	-1.458	5.545	32.873	1.543	2.255	7.466
200* and 1082	.146	.132	.039	- .503	.444	.450	.470	2.312	5.566
600 and 1082	-.036	.049	.004	.240	.090	.066	.143	.230	.073
800 and 1082	.007	.030	.001	- .009	.348	.121	.156	.210	.068

Total inconsistency is equal to the square of the bias plus the square of the standard deviation of the difference.

* Two-parameter model calibration based on 42 of 43 items for this examinee sample.

Investigating Table 4.10, it is interesting to note that less of the difference between the two methods of equating that use the two-parameter model is due to systematic bias and more is due to the standard deviation of the difference in equating. More of the difference between the two-parameter equatings and the Rasch equating is due to mean difference.

Table 4.11 contains the equating statistics for the comparisons within models across Fall 1979 examinee samples. The Rasch model is very consistent with total inconsistency ranging from .001 to .039 for all samples. By comparison, the two-parameter model total inconsistencies range from .066 to .121 for samples of 600 and 800. For the sample of 200 examinees based on 43 calibrating items, the alternate item calibration method resulted in a mean square inconsistency of 7.466, 191 times larger than the Rasch inconsistency. The item difficulty calibration method resulted in an inconsistency of 32.873, over four times larger than the alternative method inconsistency and 843 times larger than the Rasch inconsistency. On the other hand, for the sample of 200 based on 42 calibrating items, the alternate item calibration method resulted in a mean square inconsistency of 5.566, which is 142 times larger than the Rasch inconsistency. The item difficulty calibration method resulted in an inconsistency of only .450, 11½ times the Rasch inconsistency.

Consistency of equating across examinee samples is an important property of an equating procedure. However, an equating method may be consistent but result in incorrect equatings. The next section presents the results of equating where a clear-cut criterion exists.

Study 2

The results of equating the Fall 1979 examination to itself based on the three sets of examinee samples are presented in Figures 4.11 through 4.16. A correct equating would result in a line with a slope of one and an intercept of zero. It can be seen that the Rasch equatings seem in general closer to the criterion line than either of the two-parameter equatings.

Table 4.12 contains equating error statistics for the equatings shown in Figures 4.11 through 4.16. The statistics were computed based on equation (4.2). For Table 4.12, however, the summation ranges over ability estimates corresponding to true scores between approximately 25 and 97. Since the three sets of examinations were not placed on the common reference scale, it was necessary to use a common true-score range to make the comparisons meaningful. The range from chance (25) to the upper limit conveniently available (97) seemed reasonable. Therefore, the values within Table 4.12 are directly comparably to one another. However, although the values in Table 4.12 are in the same metric as those in Tables 4.10 and 4.11, they are not directly comparable in magnitude since they are based on a somewhat different underlying ability range.

Investigating Table 4.12, it can be seen that no method is uniformly best, but that the item difficulties method of calibration results in worse equating than the Rasch model for all three samples. The other two-parameter model is better only for the random samples and very much worse for the low ability - high ability samples.

Study 3

Two types of error in equating tests using the Rasch model in Study 3 are error in assigning abilities within a test and error caused

FIGURE 4.11: RASCH RANDOM HALVES

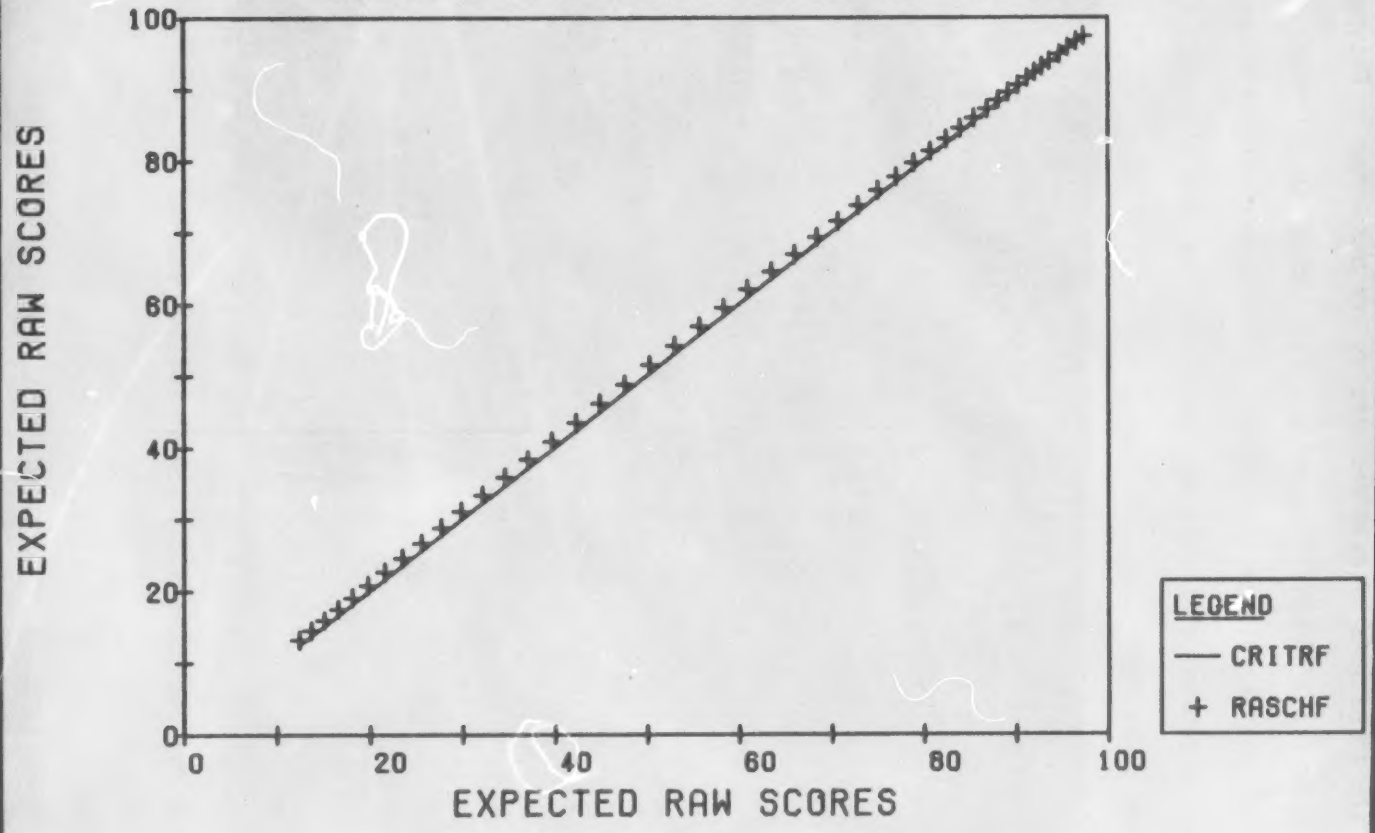


FIGURE 4.12: RASCH LOW TO HIGH HALF

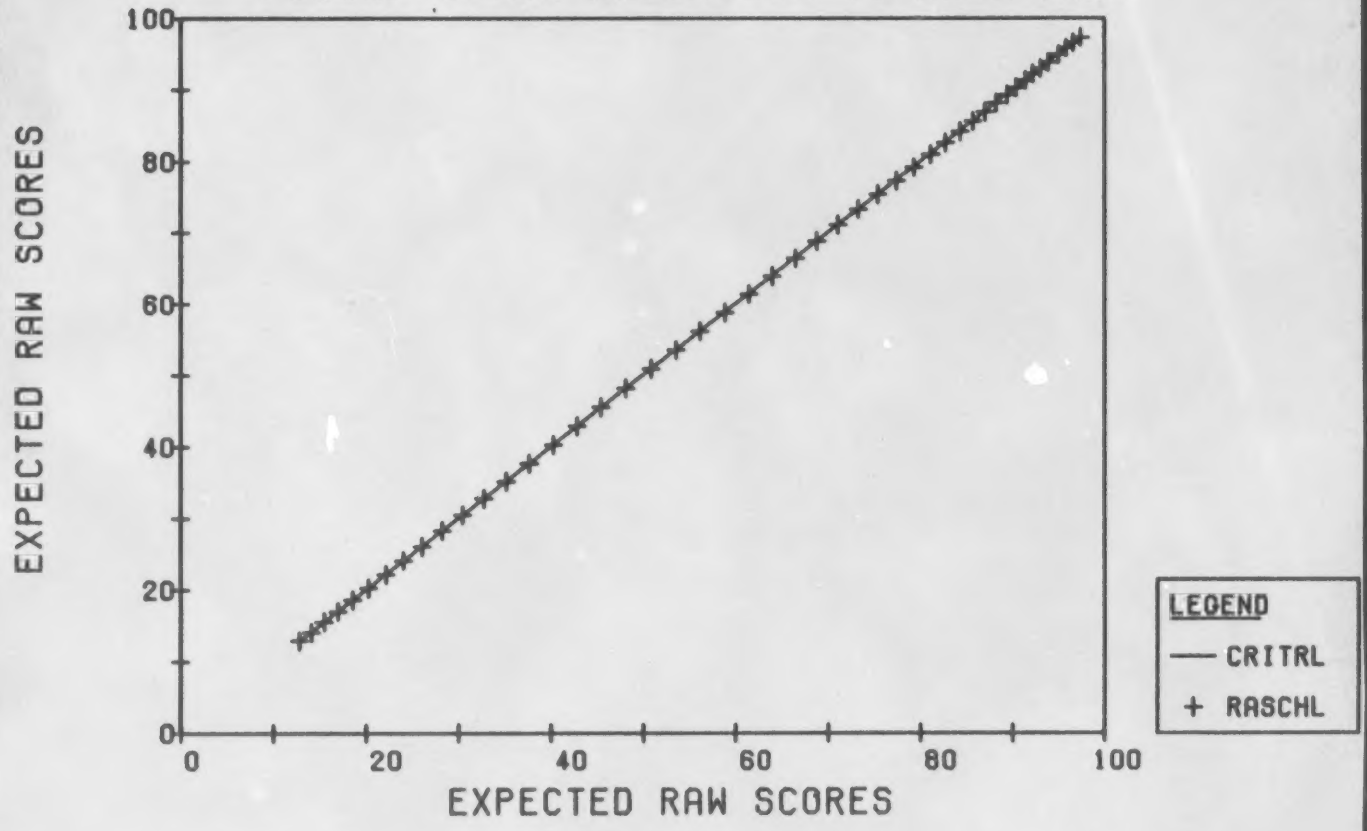


FIGURE 4.13: RASCH VERY LOW TO VERY HIGH HALF

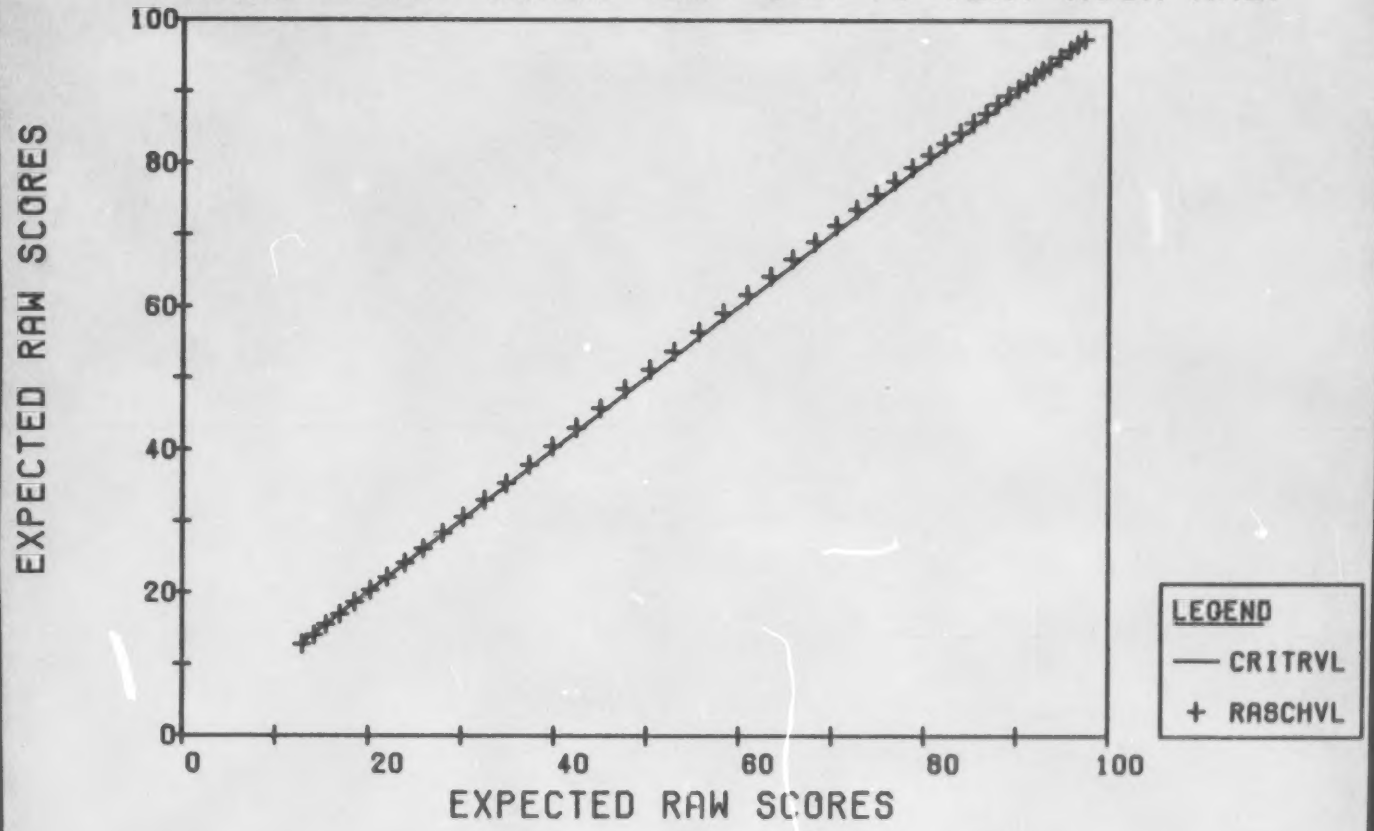


FIGURE 4.14: 2 PAR RANDOM HALVES

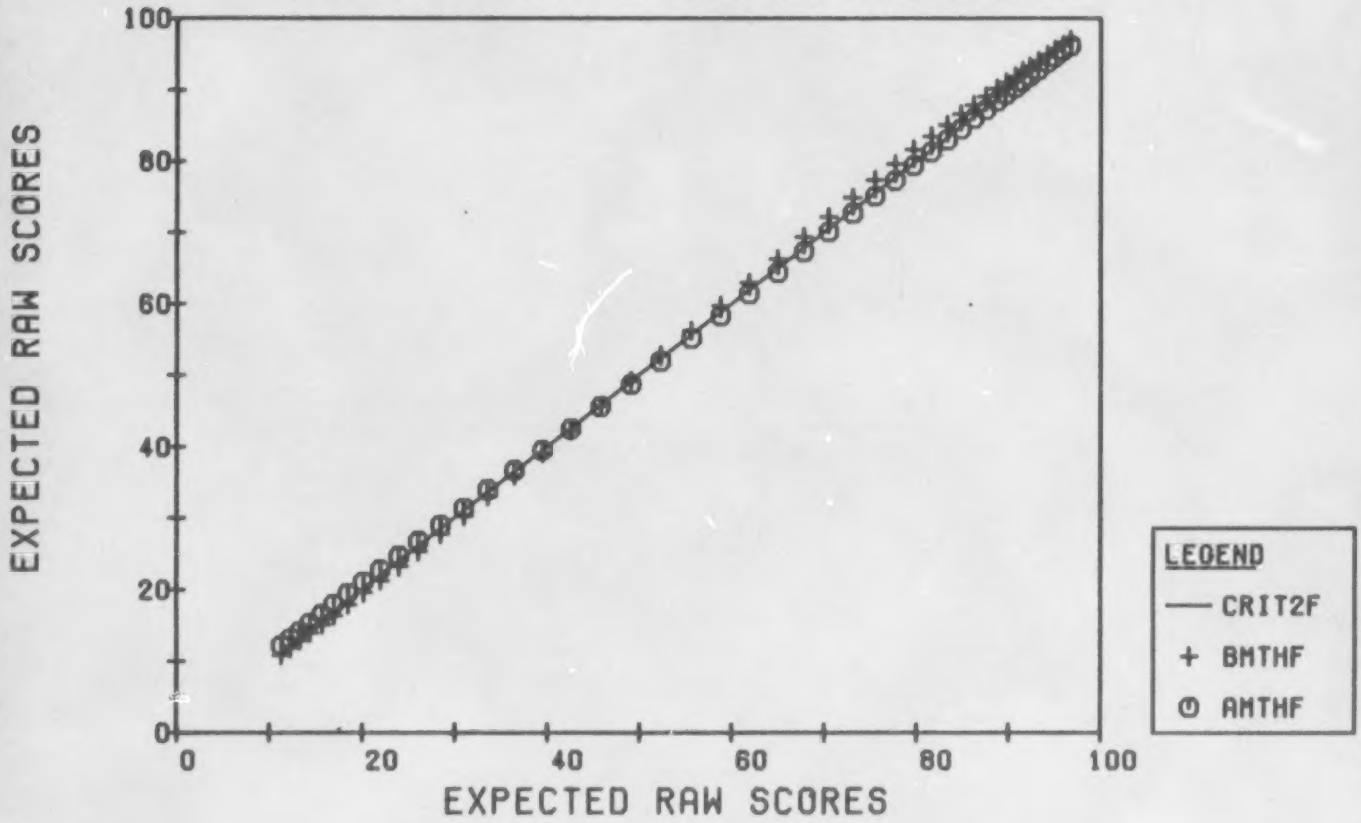


FIGURE 4.15: 2 PAR LOW TO HIGH HALF

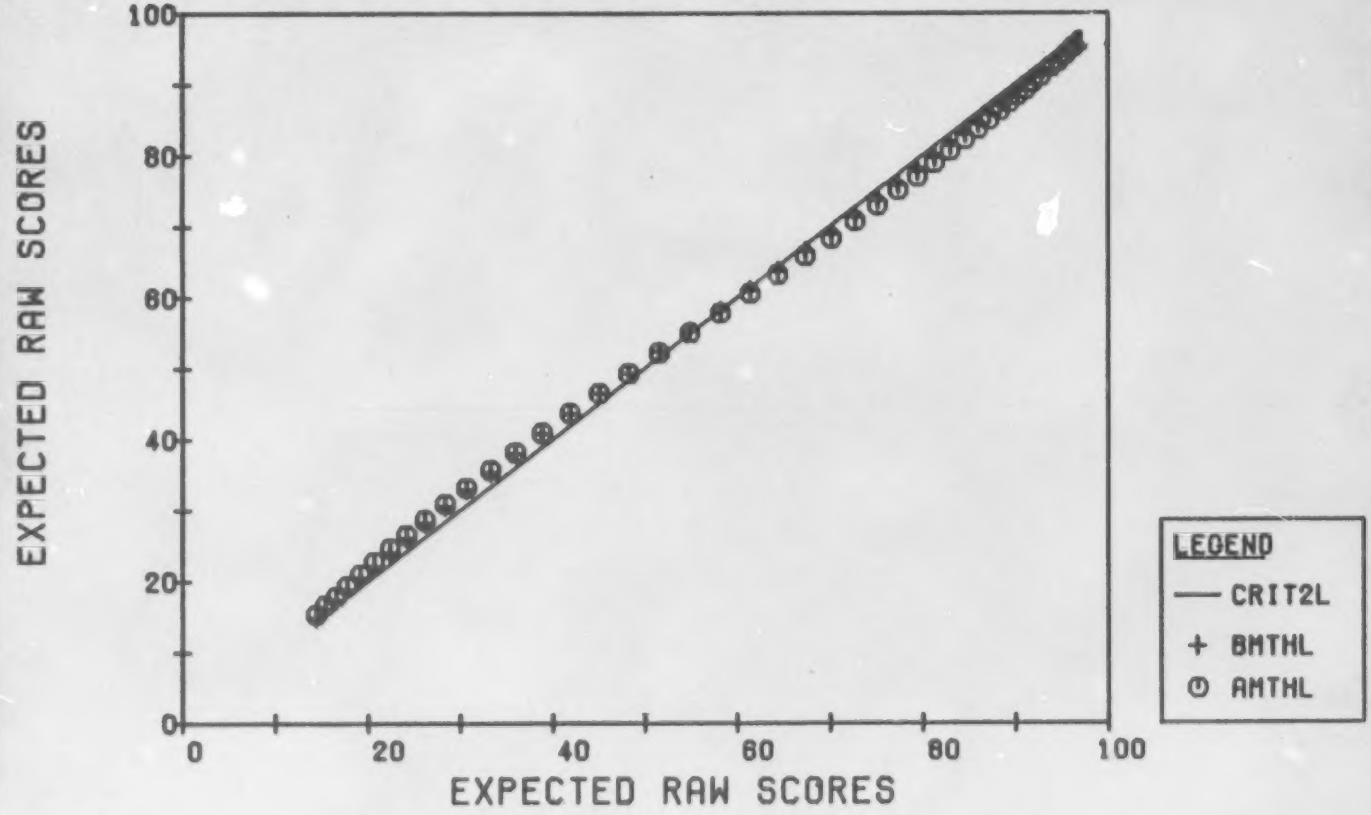


FIGURE 4.16: 2 PAR VERY LOW TO VERY HIGH HALF

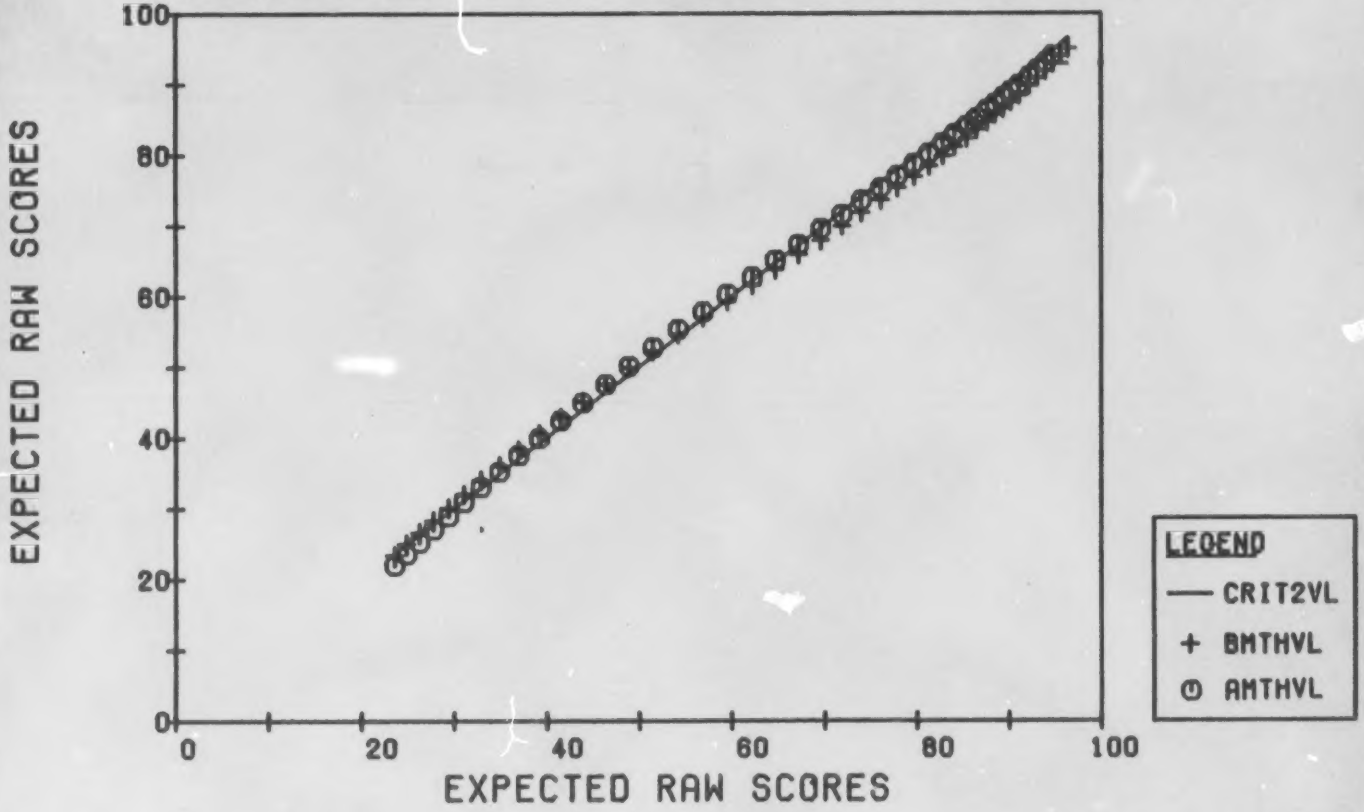


TABLE 4.12

EQUATING ERROR ACROSS MODELS

	RASCH			TWO-PARAMETER B METHOD			TWO-PARAMETER A METHOD		
Equating Samples	Bias	S.D. Diff.	Total	Bias	S.D. Diff.	Total	Bias	S.D. Diff.	Total
Random	.808	.375	.793	.525	.926	1.133	-.049	.540	.294
Low - High	-.039	.053	.004	- .174	1.320	1.773	-.400	1.723	3.129
Very Low-Very High	.484	2.90	.318	-1.205	1.558	3.879	-.454	.839	.910

Total error is equal to the square of the bias plus the square of the standard deviation of the difference.

by incorrect calibration of abilities across tests. The first source of error is investigated in Table 4.13 and the second in Table 4.14.

Table 4.13 contains Rasch ability estimates corresponding to raw scores based on the 43 item subtest selected from the Fall 1979 examination. Estimates are presented based on examinees taking the Spring 1979 and Fall 1979 examination. Estimates for the two class sections, taking the Fall 1979 examination, that have the most extreme ability estimates compared to the other 14 sections are also presented. The standard deviations of ability are taken from the Fall 1979 sample but differ from the other samples by less than .05 at every ability level. The discrepancy between the estimates based on the Spring 1979 and Fall 1979 samples varies from .00 in the middle of the test score range to .19 at a raw score of one. No discrepancy in Table 4.13 is greater than .40 standard errors.

Table 4.14 contains the mean, standard deviation, maximum negative and maximum calibration constant error (E_k) across the 16 sections for Study 3. The criteria for determining the correct calibration constant are the calibration based on all 43 items and either the scale determined by the 43 items on the Fall 1979 examinee sample or the 100 items on the Spring 1979 examinee sample. The Spring 1979 examination sample provides the more realistic criterion.

Notice from Table 4.14 that there is a systematic error in calibration constants. Using the Spring 1979 criterion, the average calibration error across sections ranges from -.068 to -.192. Even with only six of 43 items removed from calibration, there is an average error of -.119 and a maximum error of -.167. Assuming a person ability standard deviation of .67, on the average persons in all the sections would be assigned ability

TABLE 4.13

RASCH ABILITY ESTIMATES FOR STUDY 3

Raw Score	$\hat{\theta}$ Spring 1979	$\hat{\theta}$ Fall 1979	$\hat{\theta}$ Section 4 Fall 1979	$\hat{\theta}$ Section 15 Fall 1979	S.E. $\hat{\theta}$
42	3.95	3.77	4.02	3.76	1.02
41	3.23	3.07	3.30	3.05	.73
40	2.79	2.64	2.86	2.63	.60
39	2.47	2.33	2.53	2.32	.53
38	2.21	2.08	2.27	2.07	.48
37	1.99	1.87	2.05	1.86	.45
36	1.80	1.69	1.85	1.68	.42
35	1.62	1.53	1.68	1.51	.40
34	1.47	1.38	1.52	1.37	.38
33	1.32	1.24	1.37	1.23	.37
32	1.19	1.11	1.23	1.10	.36
31	1.06	.99	1.10	.98	.35
30	.93	.87	.97	.86	.34
29	.82	.76	.85	.75	.33
28	.70	.65	.74	.65	.33
27	.59	.55	.62	.54	.33
26	.48	.45	.51	.44	.32
25	.38	.35	.40	.34	.32
24	.27	.25	.30	.25	.32
23	.17	.15	.19	.15	.32
22	.06	.05	.08	.05	.32
21	-.04	-.04	-.02	-.04	.32
20	-.14	-.14	-.13	-.14	.32
19	-.25	-.24	-.24	-.24	.32
18	-.35	-.34	-.35	-.34	.32
17	-.46	-.44	-.46	-.44	.32
16	-.57	-.54	-.57	-.54	.33
15	-.68	-.65	-.69	-.64	.33
14	-.80	-.75	-.81	-.75	.34
13	-.92	-.87	-.93	-.86	.34
12	-1.05	-.98	-1.06	-.98	.35
11	-1.18	-1.11	-1.20	-1.10	.36
10	-1.32	-1.24	-1.35	-1.23	.37
9	-1.47	-1.38	-1.50	-1.37	.38
8	-1.63	-1.53	-1.67	-1.52	.40
7	-1.80	-1.69	-1.86	-1.68	.42
6	-2.00	-1.88	-2.07	-1.86	.45
5	-2.23	-2.09	-2.31	-2.07	.48
4	-2.49	-2.34	-2.59	-2.32	.53
3	-2.82	-2.66	-2.94	-2.64	.61
2	-3.26	-3.09	-3.41	-3.06	.74
1	-3.99	-3.80	-4.17	-3.77	1.02

TABLE 4.14

RASCH ABILITY ESTIMATE ERROR DUE TO CALIBRATION FOR STUDY 3

Origin of Comparison Scale

Term	# Items	# Calibrating Items	Average E_k	S.D. E_k	Maximum Negative E_k	Maximum E_k
Fall 1979	43	37	-.002	.036	-.050	.061
Fall 1979	43	32	.002	.043	-.076	.077
Fall 1979	43	27	-.002	.074	-.128	.128
Fall 1979	43	22	-.007	.096	-.175	.188
Fall 1979	43	17	-.013	.120	-.316	.154
Fall 1979	43	12	-.006	.146	-.358	.170
Fall 1979	43	7	-.023	.233	-.570	.330
Spring 1979	100	37	-.119	.036	-.167	-.056
Spring 1979	100	32	-.076	.043	-.154	-.001
Spring 1979	100	27	-.120	.074	-.246	.010
Spring 1979	100	22	-.138	.096	-.306	.057
Spring 1979	100	17	-.192	.120	-.495	-.025
Spring 1979	100	12	-.149	.146	-.501	.027
Spring 1979	100	7	-.068	.233	-.615	.285

estimates .18 standard deviation units too low and for the most discrepant section ability estimates would be .25 standard deviations too low.

Descriptive data on the examinations and examinee samples and parameter estimation information were presented in the first two sections of this chapter. Calibration statistics and their impact on test equating and ability estimation comprised the remainder of the research results. The next and final chapter summarizes the study, draws conclusions from the findings and presents recommendations based on the findings for classroom examination systems.

CHAPTER V

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

SUMMARY

Item characteristic curve models offer the potential of producing a measurement of a person's ability or achievement that does not depend on the particular sample of items in a test or on the particular examinees upon which test analysis is based. There has been much research effort devoted to ICC models using simulated or aptitude test data. Perhaps because of the apparent multidimensionality of achievement tests, there has been very little research focused on tests of academic achievement. If ICC models are to be of any use in classroom testing, more ICC research must be focused on achievement testing in general and on classroom achievement testing in particular.

The present study is designed to provide methods and results relevant to the introduction of item characteristic curve models into classroom achievement testing. The overall objective is to compare several common ICC models for item calibration and test equating in a classroom examination system.

Parameters for the one-, two- and three-parameter logistic ICC models were estimated for 100-item final examinations given over the course of four academic terms. The three-parameter model provided unacceptable lower asymptotes for this data set and was given no further

consideration. One of the four tests was chosen to provide a common scale to which items from the other examinations were calibrated using the one- and two-parameter models. Random samples of 200, 600 and 800 examinees were selected from a total sample of 1082 examinees to facilitate the investigation of the impact of examinee sample size upon item calibration and test equating. Rasch calibration and two methods of two-parameter model calibration were investigated. The calibrations based on the examinee sample of 200 and the entire set of items in the calibration section were quite divergent across methods and also diverged from the calibrations based on the other samples.

The calibration constants estimated from the Rasch model and the two methods using the two-parameter model were used to equate the examinations to a common scale. The Rasch model equatings were very consistent across all sample sizes investigated. The two-parameter methods provided less consistent equatings than those provided by the Rasch model.

Consistency of equating provides only indirect evidence of equating accuracy. To provide a criterion for equating, three separate sets of examinee samples were selected from the 1082 examinees in the largest total sample. The three sets represented random, high-low and very high-very low examinee ability differences. Although none of the three methods was uniformly best, the Rasch model provided the most acceptable equating in general.

The stability of Rasch calibrating links and hence ability estimation based on examinee class sections was investigated for a 43 item test taken from a total test of 100 items. Calibration links were calculated for subsets of the 43 items ranging in size from 7 to 37 in increments of five items. Even with 37 items in the calibrating link, it was

found that there was an overall bias in ability estimation of approximately 1/5 of a standard deviation of ability. For the most extreme group, there was a bias of 1/4 of a standard deviation using 37 items.

CONCLUSIONS AND DISCUSSION

The previous section provided a brief review of the problem, methods and results. This section presents some conclusions that can be drawn from the results and is followed by a final section containing recommendations for further research.

Parameter Estimation

There are two reasonable explanations for the inadequacy of lower asymptote estimates for the three-parameter model. First, for the four examinations studied, there is very little useful information available for estimating the lower asymptotes. In order to accurately estimate the probability of very low ability examinees answering a question correctly, it is necessary to have some very low ability examinees in the sample. The examinations had means and standard deviations in the neighborhood of 69 and 13, respectively. Obviously there are not very many examinees at all close to a chance score of 25. The second explanation is that the problem of adequately estimating lower asymptotes has not been solved. However, newer versions of LOGIST are using maximum likelihood estimation procedures that should give better estimates.

Unless there are very low scoring persons in an examinee sample, it is probably futile to attempt to estimate parameters for the three parameter model. For common classroom achievement tests this is unlikely. Several alternatives are available. One is to give the test as a pretest to a group of examinees and combine their responses with those of an ordi-

nary group. A second alternative is to set the lower asymptotes for all items at a reasonable value, for example .05 below chance for well-written items. Further research is needed to identify the models for which parameters can be estimated well under different conditions.

Calibration

Problems still remain with calibration methods for the two-parameter model. One undesirable property of both methods used in this study is that it is necessary to remove some of the items from the calibrating sample before determining a calibration equation. The removal of these outliers is a subjective and somewhat arbitrary task. The removal or retention of even one item can make a meaningful difference in the calibration equation. Lord (personal communication, April 1980) is pursuing a promising technique using robust estimators. He is attempting to estimate the parameters of calibration equations by using estimates that are little affected by outliers and yet have high efficiency in a variety of situations.

Neither of the two methods considered in this study for calibrating two-parameter estimates provided more consistent results than the other across different size samples. It should be noted that goodness of calibration depends on the degree of linearity between item difficulty estimates for common items across tests. The correlations for the tests in this study were not impressive. For the total examinee samples they ranged from .57 to .87 (Table 4.8). The higher coefficients were obtained only after outlying items had been discarded. The lack of a strong linear relationship may limit the accuracy of any calibration technique.

The Rasch calibrations in general displayed remarkable consistency across types and sizes of examinee samples taken from a single test

administration. However, a very disturbing fact became apparent in Study 3. When calibration constants based on different subsets of items and different sections of examinees were compared to those obtained from the total sample, they displayed reasonable congruence. But when the more realistic criterion of similarity of calibration across examinations was used, the calibrations were incorrect but reasonably consistent with one another. In other words, there was a systematic error in calibration across examinations. This is a particularly disturbing result since the invariance of item characteristic curve statistics and ability estimates clearly depends on a stable scale. Scale drift has also been found using the three-parameter model at the Educational Testing Service (Marco, personal communication, April 1980). It looks as though scale drift may be a common phenomenon with latent trait models. It is certainly a possibility that must be recognized and investigated in any testing system to which ICC models are being applied.

Test Equating and Ability Estimation

True-score test equating depends upon parameter estimates and item calibration, so it is not surprising to find that the equating results mirror the calibration findings. The Rasch equatings across the various sample sizes were extremely consistent. Both of the two-parameter calibration methods resulted in reasonably consistent equatings for samples of 600 and 800 compared to 1082 and very divergent equatings for samples of 200 when all 43 items were used for calibration.

The consistency of the Rasch model, even with examinee samples of 200 and all items used for calibration, coupled with its generally superior equatings of a test with itself, provides support for its adoption. However, two caveats need to be made. The first is that in Study 3 the

Rasch calibrations were consistent but incorrect. The second caveat is that equating a test to itself seems somewhat biased in favor of a model with less parameters (Marco, Peterson and Stewart, 1980). True-score equating is accomplished by setting up a correspondence between the sums of the ICC's taken from two tests at a number of different ability points. When a test is equated to itself the proper equating will occur if the sums of the ICC's are the same at every point. Since the items are the same across the tests when a test is equated to itself, any consistent assignment of parameter estimates across the two samples will result in a correct equating. The two-parameter model has two parameter estimates that may differ across samples for each item. The one-parameter model only has one opportunity to differ for a given item. Therefore, the criterion of equating a test to itself seems to favor the one-parameter model which has only half as many opportunities to produce incorrect results.

The equatings of the test to itself were somewhat surprising in that the amount of difference between ability groups appeared to have no consistent relationship to the goodness of equating. However, this is consistent with the finding of Marco, Peterson and Stewart (1980) that types of examinee samples have little effect on equating if the anchor test is similar in content and difficulty to the tests being equated.

In summary, the Rasch model, despite some concern that one criterion may be biased in its favor, seems to be providing better achievement test equatings than the two-parameter model in this study. It is clearly more consistent than the two-parameter model, especially for smaller sample sizes. Given the greater ease of using and interpreting the Rasch model and its lower cost, the Rasch model seems to most warrant further study in the Communication 100 testing system.

RECOMMENDATIONS

In general the problem of scale drift associated with item characteristic curve models needs to be investigated more fully. Any design for implementing ICC methods in a testing system should include a systematic study of changes in item parameter estimates over time. Items with unstable estimates can be identified and removed from item pools. Studies should be designed to investigate and control the amount of scale drift over time.

In the context of the Communication 100 testing system and others like it, there are three specific recommendations. First, continue investigating the Rasch model and calibrating items to the Spring 1979 scale. The Rasch model seems to be working well enough to warrant further study.

The second recommendation is to work to improve the results of the Rasch model by evaluating items currently in the pool as well as new items using criteria relevant to the Rasch model and using these criteria to select items for retention or replacement. The criteria should include comparisons of items on item tests of fit supplied by BICAL and on stability of difficulty estimates over time.

The third recommendation is to explicitly study and attempt to control scale drift. Methods of controlling scale drift include improving the fit of items to the Rasch model and removal of items displaying unstable difficulty estimates from the item pool and from calibration sections before calibration. The effect of scale drift on test equating can be determined by administering a test previously given, equating the test in the normal way and then comparing the new equating to the original equating.

Classroom achievement tests present some unique problems to those seeking the advantages of ICC models. They tend to be less unidimensional than standardized aptitude tests and their dimensionality may change over time due to instructor or curriculum changes. Typically, they are of relatively low difficulty for the examinees, so lower asymptotes of the three-parameter model may be hard to estimate. Examinee sample sizes are typically much smaller than for standardized tests and item quality is usually lower. However, the potential advantages of ICC models are great. If ICC models can be made to work properly, estimates of ability or achievement will not be dependent on the particular sample of items in an examination nor upon the particular students in a classroom. The advantages of ICC models are sufficiently great and the results of attempts to apply ICC models are sufficiently positive to warrant further research on and applications of ICC models.

APPENDIX